# Werkzeuge für Rechtsdatenbanken

## Über computerlinguistische Verfahren zur Untersuchung, Speicherung und Kommunikation rechtlichen Wissens

Inaugural-Dissertation
zur Erlangung des Doktorgrades der Philosophie
an der Ludwig-Maximilian-Universität
München

vorgelegt von

Leonhard Voltmer

aus München

veröffentlicht im Jahre 2005

| | |
|---|---|
| Referent: | Prof. Dr. Wulf Oesterreicher |
| Korreferent: | Prof. Dr. Klaus U. Schulz |
| Tag der mündlichen Prüfung: | 27.01.2005 |

# Abkürzungsverzeichnis

| | |
|---|---|
| a.a.O. | an anderem Ort |
| Art. | Artikel |
| BGH | Bundesgerichtshof |
| BISTRO | Bozner Informationssystem für Rechtsterminologie |
| CC | Codice Civile (italienisches Zivilgesetzbuch) |
| CES | Corpus Encoding Standard |
| CSS | Cascading Style Sheets |
| DDC | Dewey Decimal Classification |
| DF | Dokumentfrequenz |
| et al. | und andere |
| EU | Europäische Union |
| EURAC | Europäische Akademie Bozen/Bolzano |
| evtl. | eventuell |
| ff. | fortfolgende |
| FOB | Formating Object |
| Fußn. | Fußnote(n) |
| GVG | Gerichtsverfassungsgesetz |
| HTML | Hypertext Markup Langauage |
| IR | Information Retrieval |
| ISBN | Internationale Standardbuchnummer |
| ISO | International Standards Organization |
| LMBG | Lebensmittel- und Bedarfsgegenständegesetz |
| MAB | Maschinelles Austauschformat für Bibliotheken |
| MI | Mutual Information |
| MIRIS | Minority Rights Information System |
| NLP | Natural Language Processing |
| NN | nächster Nachbar |
| Nr. | Nummer |
| PDA | Personal Digital Assistant |
| PDF | Portable Document Format |
| POS | Part of Speech |
| s.o. | siehe oben |
| Schuldrecht AT | Schuldrecht allgemeiner Teil |
| Schuldrecht BT | Schuldrecht besonderer Teil |
| SGML | Standard Generalized Markup Language |
| SQL | structured query language |
| SVG | Scalable Vector Graphics |
| TC | Termkandidat |
| TE | Termextraktion |
| TEI | Text Encoding Initiative |
| TF | Termfrequenz |
| UDC | Universal Decimal Classification |
| URL | Uniform Resource Locator |
| usw. | und so weiter |
| UWG | Gesetz über den unlauteren Wettbewerb |
| v. | von/vom |
| WAP | Wireless Application Protocol |
| XCES | Corpus Encoding Standard in XML |
| XML | Extended Markup Language |
| z.T. | zum Teil |

Inhaltsverzeichnis

Verzeichnis aller Grafiken und Tabellen

# Einleitung: Schriftlichkeit und Recht

Recht und die Rechtswissenschaft manifestieren sich vor allem über Schrift. Die Schriftlichkeit ist ein wesentlicher Bestandteil des Rechts geworden. Die Untersuchung der Schrift mit den modernsten computerlinguistischen Methoden verspricht daher reiche wissenschaftliche Ernte. In dieser Arbeit werden auf einigen besonders fruchtbaren Feldern Methoden diskutiert und ihre Umsetzung in Werkzeuge evaluiert.

## 1. Bedeutung der Schriftlichkeit für das Recht

Recht könnte man als die Gesamtheit der Normen für menschliches Verhalten in der Gesellschaft definieren. Zunächst hat Recht also keinerlei zwingenden Bezug zur Schriftlichkeit und funktionierte bis 2100 Jahre vor unserer Zeitrechnung erste Rechtsurkunden und Codices erstellt werden, deren Bedeutung für das Rechtssystem (Vollständigkeit, Systematik, Rechtskraft) noch diskutiert wird.[1]

Auch heute werden die Normen für menschliches Verhalten in den meisten Staaten Afrikas und Asiens, vor allem je weniger stark sie kolonialem Einfluss unterlagen, von einer mündlichen Rechtstradition gesetzt und erhalten. Die Verschriftung von Normen auch in diesen Ländern darf über ihre Herkunft und Bedeutung als nicht hinwegtäuschen. Auch das Zwölftafelgesetz, die *Sharia* und die Gesetzesbestimmungen des Alten Testaments sind keine neuen Gesetze, sondern die schriftliche Sammlung bestehenden Gewohnheitsrechts.[2]

Im europäischen Mittelalter wandelte sich dann das Verständnis der Rechtsquelle vom selten verschrifteten Gewohnheitsrecht zum autoritativ erlassenen Rechtsgebot. Zusammen mit der Erfindung des Buchdrucks mit beweglichen Lettern und einer Fülle weiterer Faktoren nahm die Verbreitung schriftlicher Gesetzestexte zu. Adressaten waren aber vorrangig die gebildeten Schichten, bis die Alphabetisierungsrate durch die allgemeine Schulpflicht zu steigen begann.[3] Aber auch die gebildeten Schichten waren von der Kultur der Mündlichkeit beeinflusst: Schriftliche Urkunden waren über lange Zeit hinweg nicht konstitutiv sondern dokumentarisch. Außerdem gab es keine allgemein anerkannten Auslegungsregeln, so dass für die Geltung keinerlei Unterschied nach der örtlichen, sachlichen, zeitlichen oder hierarchischen Herkunft von Rechtsdokumenten gemacht wurde.[4]

Mit dem Ausbau der Schriftlichkeit im Recht geht ein zweifacher Ausbau der Fachsprache einher. Zum einen können immer mehr kommunikative Situationen mit der Rechtssprache bestritten werden (extensiver Ausbau), zum anderen werden die sprachlichen Ausdrucksmittel im Vergleich mit der Umgangssprache oder anderen Sprachsituationen immer spezieller (intensiver Ausbau). In der Beziehung konzeptionell mündlicher zu konzeptionell schriftlicher Kommunikationssituationen nach Koch/Oesterreicher[5] markieren Gesetzestexte den Pol der Schriftlichkeit. Andere rechtliche Kommunikationssituationen wie Gerichtsurteile oder rechtswissenschaftliche Aufsätze sind ebenfalls ganz auf der Seite der konzeptionellen Schriftlichkeit zu vermuten. Damit gelten die typischen Sprachmerkmale konzeptioneller Schriftlichkeit insbesondere für rechtliche Kommunikationssituationen.

---

[1] Kienast B. (1994), Die Altorientalischen Codices zwischen Mündlichkeit und Schriftlichkeit, S. 13-26 in: Gehrke H.-J. (Hrsg.)(1994), Rechtskodifizierung und soziale Normen im interkulturellen Vergleich, Gunter Narr Verlag, Tübingen, 1994, S. 17.

[2] Pegoraro L., Reposo A. (1993), Le fonti del diritto negli ordinamenti contemporanei, Monduzzi Editore, Bologna, 1993, S. 18.

[3] Holzborn T. (2003), Die Geschichte der Gesetzespublikation, Tenea Verlag, Berlin 2003, S. 132 f.

[4] Patzold S. (2000), Konflikte im Kloster, Matthiessen Husum 2000, S. 342 spricht von der „Rechtsmentalität".

[5] Koch P., Oesterreicher W. (1994) Schriftlichkeit und Sprache, S. 587-604 in: Hartmut Günther., Otto Ludwig (Hrsg.) (1994): Schrift und Schriftlichkeit. Ein interdisziplinäre Handbuch internationaler Forschung 1. Halbband, Berlin/New York 1994.

Nach Sieber[6] sind dies u.a. die folgenden sprachlichen Mittel: schwierigere, differenziertere, längere, variationsreichere Lexik und Syntax, weniger Partikeln, mehr Präteritum und synthetischer Konjunktiv II, mehr Information im Verhältnis zur Textlänge. Dazu kommen folgende pragmatische Bedingungen: monologisch, tendenziell öffentlich mit mehr und anonymen Kommunikationspartnern, geplant, mit langer Produktions- und Rezeptionszeit, zerdehnte Kommunikationssituation (Zeit- und Ortsversetztheit von Produktion und Rezeption), verselbständigt, nur visueller Zeichenkanal und zumeist nur eine Zeichenart. Die kommunikative Grundhaltung der Schriftlichkeit hat folgende Merkmale: An Distanz orientiert, literales Wissen und Verstehen wichtig, Integration von Sachwissen in einen Wissensdiskurs angestrebt, textuelle Organisation als Sprachwerk wichtig und in der Rezeption wird das Wort beim Wort genommen (*literal meaning*). Diese Besonderheiten der rechtlichen Fachsprache lassen sich beliebig belegen. Allein der letztgenannte Punkt, das Haften am Wortlaut, beherrscht die frühe Entwicklung in der Schriftlichkeit des Rechts[7] und in abgemilderter Form vor allem die Strafrechtslehre bis heute.

Die besonderen sprachlichen Mittel, sprachlichen Bedingungen und die kommunikative Grundhaltung bei der Verwendung von Schrift im Recht beschreiben nicht nur eine von vielen möglichen Formen rechtlicher Kommunikation. Praktisch die gesamte relevante rechtliche Kommunikation läuft über die Schrift.

Rechtsstaaten verlangen die Schriftlichkeit im Gesetzgebungsverfahren.[8] Selbst die ‚Verkündung' eines Gesetzes, meist letzter Verfahrensschritt, muss schriftlich im offiziellen Gesetzesblatt erfolgen, obwohl der Ausdruck darauf hinweist, dass die Veröffentlichung einst mündlich erfolgte. Die Schriftsprache benötigt keine ‚Übersetzung' mehr in die gesprochene Sprache, sie hat sich als alleiniges Konzept und Medium des Rechts emanzipiert.[9] Im Gegenteil bedarf die mündliche Verhandlung der Übersetzung in die Schriftlichkeit (Protokollierung), um rechtlich relevant zu sein.[10]

Die Vormacht des Geschriebenen wird noch dadurch verstärkt, dass die juristische Auslegung sich von der philologisch-historischen dadurch unterscheidet, dass sie nicht den vom Urheber gemeinten Sinn, sondern einen objektiven, dem Text immanenten Sinn zu erfassen versucht.[11] Rechtstexte entstehen zwar wie alle Texte aus einem bestimmten Zusammenhang heraus, sie sind aber auf

---

[6] Sieber P. (1998), Parlando in Texten. Zur Veränderung kommunikativer Grundmuster in der Schriftlichkeit, S. 182-188 in: Henn H., Sitta H., Wiegand H.E. (Hrsg.) (1998), Reihe Germanistische Linguistik 191, Max Niemeyer Verlag Tübingen 1998, m. w. Nachw.

[7] Honsell, H. (2001): Naturrecht und Positivismus im Spiegel der Geschichte, S. 593-602 in: Festschrift Koppensteiner, Orac Verlag Wien 2001, S. 596: „In primitiven Rechtsordnungen kommt es nur auf den Wortlaut der gesetzlichen Ge- und Verbote an, nicht auf ihren Sinn". Auch im Mittelalter bestimmte allein der Wortlaut den Sinn.

[8] Falsch ist hingegen die Aussage „In modern society all legal provisions can be found in writing.", Heutger V. (2000), Law and Language in the European Union, S. 4 in: Global Jurist Topics, Vol. 3, Issue 1, 2000, Article 3. Wenn mit *legal provisions* Rechtsbestimmungen gemeint sind, dann stimmt die Aussage für die meisten Rechtsordnungen nicht, weil auch Gewohnheitsrecht wie ungeschriebene Handelsbräuche Rechtsbestimmungen sind. Wenn *legal provisions* Gesetze im technischen Sinn gemeint sein sollten, dann ist die Aussage, dass diese auch in Schriftform zu finden sind, selbstverständlich und sehr schwach, denn das Gesetzgebungsverfahren verlangt die Schriftform und verweigert mündlichen Gesetzen die Rechtskraft.

[9] Brockmeier spricht von der „kulturellen Hegemonie eines Mediums", die zumindest für das Recht belegt werden kann. Brockmeier J. (1999), Die Welt als Bibliothek, Colloquium vom 15.7.1999, http://www.fu-berlin.de/postmoderne-psych/berichte3/brockmeier.htm: 13.11.2003.

[10] Vergl. §§ 159-165 ZPO, insbes. § 165 Abs. 1 S. 1 ZPO: „Die Beachtung der für die Verhandlung vorgeschriebenen Förmlichkeiten kann nur durch das Protokoll bewiesen werden." Vergl. §§ 168-168b StPO und insbes. § 274 S. 1 StGB: „Die Beobachtung der für die Hauptverhandlung vorgeschriebenen Förmlichkeiten kann nur durch das Protokoll bewiesen werden."

[11] Sog. ‚objektive' Auslegungstheorie. Busse D. (2002), Juristische Auslegungstätigkeit in linguistischer Hinsicht, S. 136-162 in: Haß-Zumkehr U. (Hrsg.) (2002), Sprache und Recht, Jahrbuch 2001 des Instituts für deutsche Sprache, de Gruyter Verlag Berlin 2002, S. 139 mit Verweis auf die Methodenlehre der Rechtswissenschaft von Larenz.

Abstraktion von der konkreten Anwendungssituation hin konzipiert.[12] Der Text soll in allen Situationen Sinn haben, auf die im Text als relevante Merkmale Bezug genommen wird. Der Sinn des Geschriebenen liegt also jedermann jederzeit vor Augen.[13]

Die Schriftlichkeit steht als Konzept im Zentrum des Rechts und der Rechtswissenschaft. Mit Hilfe der Computerlinguistik kann die Schrift durch neue Methoden untersucht werden.[14]

## 2. Bedeutung der Computerlinguistik für rechtliche Schriften

Computerlinguistik ist der Wissenschaftszweig zwischen Informatik und Linguistik, bei dem es um die maschinelle Verarbeitung natürlicher Sprachen geht.[15] Die Computerlinguistik bedient sich wie die Verarbeitung natürlicher Sprache[16] symbolischer, subsymbolischer und stochastisch-statistischer Methoden. Symbolische Methoden werden in erster Linie bei der Wissensrepräsentation angewandt, bei der durch die Anwendung von Logik auf das mit formalen Sprachen dargestellte Wissen das Ziel künstlicher Intelligenz erreicht werden soll. Subsymbolische Methoden haben ebenfalls die Weiterverarbeitung des sprachlichen Wissens zum Ziel, verwenden dazu aber beispielsweise neuronale Netze. Am weitesten verbreitet sind wohl stochastische und statistische Methoden, die Aussagen über die Wahrscheinlichkeit bestimmter in der Sprache beobachteter Phänomene machen.[17] Hier wird also auf einen kausalen Zusammenhang oder den Beweis für ein Phänomen ganz verzichtet.

Wie stark sich die Computerlinguistik mit dem Recht auseinandergesetzt hat, ist schwer einzuschätzen. Eine jüngst erschienene internationale Bibliografie zu Recht und Sprache[18] verzeichnet nur 23 Titel im Bereich Computerlinguistik bis 1994 und acht Titel im getrennt geführten Bereich Korpuslinguistik bis 1999. Diese Bibliographie bietet keinen aktuellen Überblick über das Gebiet, sie zeigt aber zumindest, dass der Einsatz computerlinguistischer Methoden zur Untersuchung schriftlichen Rechts ein neueres Spezialgebiet ist.

Unglücklicherweise werden die beiden Bereiche ‚Informatik für das Recht' und ‚rechtliche Qualifikation von Computertechniken' vermischt. Das Internationale Rechtsinformatik Symposion an der Universität Salzburg (IRIS), die „größte und bedeutendste wissenschaftliche Tagung in Österreich und Mitteleuropa auf dem Gebiet der Rechtsinformatik"[19], behandelt sowohl Anwaltssoftware und Rechtsdatenbanken wie Urheberrecht und Datenschutz. Diese Vermischung ist nicht die Ausnahme, sondern die Regel.[20] Die thematische Vermengung der beiden in Methoden und Gegenstand

---

[12] Das ist auch bei Gerichtsurteilen der Fall. Dem Urteil liegt zwar ein konkreter Lebenssachverhalt zu Grunde, der Adressat eines Urteils ist aber nicht nur der Kläger oder Angeklagte, sondern ebenso deren Rechtsvertreter, die nächst höhere Instanz, die Rechtwissenschaft und spätere Gerichte, denen ein Fall mit ähnlichen Merkmalen vorliegt.

[13] „In contrast to orality, writing offers [...] the apparent advantages of physical duration, [...] precision, and detail." Glenn H. P. (2000), Legal Traditions of the World, Oxford University Press New York 2000, S. 8.

[14] Schriftlichkeit ist hingegen keine Voraussetzung für computerlinguistische Untersuchungen, denn auch gesprochene Sprache kann, mit etwas mehr Aufwand, computerlinguistisch untersucht werden.

[15] Die Computerlinguistik (CL, *computational linguistics*) ist so alt wie der Computer selbst. Carstensen K.-U., Ebert C., Endriss C., Jekat S., Klabunde R., Langer H. (Hrsgg.) (2004), Computerlinguistik und Sprachtechnologie – Eine Einführung, Spektrum Verlag Heidelberg 2004.

[16] Englisch: *natural language processing* (NLP).

[17] Stochastische und statistische Methoden können auf symbolische und subsymbolische Zeichen angewandt werden. In der Praxis werden die Ansätze oft kombiniert.

[18] Bungarten T., Engberg J. (Hrsg.) (2003), Recht und Sprache - eine internationale Bibliographie in juristischer und linguistischer Fachsystematik, in: Hamburger Arbeiten zur Fachsprachenforschung 05, Attikon-Verlag Tostedt 2003.

[19] http://www.univie.ac.at/RI/IRIS2004 : 30.9.2004.

[20] Vergleiche die Onlinepublikationen BILETA unter http://www.bileta.ac.uk/pubs.html : 30.9.2004; *International Review of Law, Computers & Technology* des Verlags *Carfax Publishing Company,* ISSN 1360-0869 print /ISSN 1364-6885 online; *International Journal of Law and Information Technology*, Oxford University Press, Print ISSN: 0967-0769, Online ISSN: 1464-3693; *Information and Communications Technology*, Carfax Publishing Company, ISSN

unterschiedlichen Wissenschaftszweige ist aus vielerlei Gründen überaus problematisch. Im Zusammenhang dieser Arbeit erschwert diese Vermengung eine Standortbestimmung der auf Recht angewandten computerlinguistischen Methoden.

Bei aller Unsicherheit einer solchen Aussage kann man es wagen, die Anwendung computerlinguistischer Methoden auf juristische Inhalte als spärlich, punktuell und in den Anfängen bezeichnen. Dies erklärt sich durch eine Reihe von verzögernden Faktoren. Zunächst seien die ‚Berührungsängste' und ‚Abstoßungsreaktionen' genannt, wenn sich rechtliche Laien im Kreis von Juristen über rechtswissenschaftliche Inhalte äußern sollen. Das Recht ist traditionell auf Autorität ausgerichtet und konnte sich der Revolution der Naturwissenschaften im letzten Jahrhundert praktisch völlig entziehen. Die Rechtswissenschaft duldet neben sich nur Hilfswissenschaften, die bei der Subsumtion dienlich sein mögen, aber keinen Einfluss auf das eigentliche Verfahren der Rechtsfindung und -schaffung haben. Das Argument dafür ist die Kluft zwischen Sein und Sollen: Den Naturwissenschaften bleibt das Sein, der Rechtswissenschaft das Sollen vorbehalten.

Ein handfesterer Verzögerungsfaktor ist sicherlich die vertikale und horizontale Pluralität des Rechts. Es gibt nicht nur ein Recht, sondern verschiedene Rechtssysteme manchmal in gleichen, manchmal verschiedenen Sprachen, die sich in manchmal gleiche, manchmal verschiedene Rechtszweige unterteilen. Insgesamt steht man vor einer Vielzahl ineinander verschachtelter und miteinander verbundener, sich teilweise widersprechender Regelungsbereiche. Hinzu kommt die Komplexität des Rechts (eine Folge der Schriftlichkeit und der damit möglichen Ausdifferenzierung!), durch die kaum ein Ergebnis als einfach, unstrittig und dauerhaft richtig bezeichnet werden kann. Eine solche Ausgangssituation erschwert die Evaluierung.

Andererseits müssten Computer umso hilfreicher sein, je komplexer die zu bewältigende Aufgabe ist. Die computerlinguistischen Möglichkeiten entwickeln sich so rasch wie die Informatik und einzelne Teilaufgaben der Computerlinguistik (rechnergestützte Übersetzung auf Wortebene, Sprachenidentifizierung einsprachiger Texte hinreichender Länge) können bereits jetzt als gelöst betrachtet werden.[21] Daher werden in dieser Arbeit Methoden in den aktuellen Kernbereichen der computerlinguistischen Forschung auf rechtliche Inhalte angewandt. Die meisten Untersuchungsmethoden sind in anderem Zusammenhang bereits in Verwendung und werden hier übertragen und an die besondere Situation angepasst.

### 3. Welche Methoden und Werkzeuge für Rechtsdatenbanken?

Wenn in Rechtstexten Wissen über das Recht enthalten ist, dann müsste in vielen Texten viel Recht zu finden sein. Die Korpuslinguistik bearbeitet linguistische Fragestellungen anhand großer Textmengen, die ausgewählt, erworben, aufbereitet (annotiert) und dann durchsucht/bearbeitet werden.[22] Die fünf Kapitel dieser Arbeit widmen sich jeweils einem wichtigen Thema, zu dem Methoden vorgestellt, in Werkzeugen implementiert und diskutiert werden.

---

1360-0834; *Journal of Law and Information Science*, University of Tasmania, Law Faculty, ISSN 0729-1485; die Zeitschrift *Computer und Recht* des Schmidt Verlags Köln und die *Computer Law Review International* des selben Verlags. Diese Vermischung einer Anwendung der Computerlinguistik mit einem Bereich des Rechts wird als ‚selbstverständlich' empfunden: „Nach ‚Saarbrücker Art' fragt die Rechtsinformatik, wie der Jurist die ihm aufgetragenen Aufgaben mit Hilfe von EDV-Instrumenten besser als bisher erfüllen kann. Selbstverständlich richtet sich dabei der Blick auch auf die Rechtsprobleme, die durch die neuen EDV-Techniken aufgeworfen werden." http://rechtsinformatik.jura.uni-sb.de : 30.9.2004, Institut für Rechtsinformatik der Universität des Saarlandes. Nach meiner Auffassung überwiegt in der Rechtsinformatik die rechtswissenschaftliche Fragestellung und Methodik bei Weitem. Juristen betrachten Rechtsinformatik also eher als Teilgebiet des Rechts.

[21] Langer, S. (2002): Grenzen der Sprachenidentifizierung, S. 99-106 in: Tagungsband KONVENS 2002, Saarbrücken, S. 99, http://konvens2002.dfki.de/cd/pdf/19V-langer.pdf : 27.9.2004.

[22] McEnery T. (2003), Corpus Linguistics, S. 448-463 in: Mitkov R. (Hrsg.) (2003), The Oxford Handbook of computational Linguistics, Oxford University Press Oxford 2003.

Im ersten Kapitel (**Dokumentklassifikation**) wird eine Methode vorgestellt, mit der gezielt Rechtstexte aus dem Internet ausgewählt, akquiriert und geordnet in ein Korpus abgelegt werden können. Auch hier sollen die Voraussetzungen so gering wie möglich gehalten werden, damit möglichst breiter Gebrauch von der Methode gemacht werden kann.

Die Einteilung des Rechts in einzelne Fachgebiete[23] hat weitreichende Folgen. Sowohl Texte wie Rechtskonzepte erlangen ihre spezielle Bedeutung durch ihr Fachgebiet. Das zweite Kapitel (**Fachgebietsklassifikation**) gibt einen Überblick über die Problematik der Fachgebietseinteilung und stellt zwei automatische Fachgebietserkenner vor, die diese Spezialaufgabe besser lösen als die in Kapitel 1 vorgestellte allgemeine Dokumentklassifikation.

Eine große Veränderung erfährt die Rechtsterminologie und -terminografie durch den Übergang von der physischen zur elektronischen Schrift. Damit muss nicht mehr eine Darstellungsweise allen Anforderungen gerecht werden, sondern die Darstellung kann dynamisch an die Umstände angepasst werden. Im dritten Kapitel (**Dynamische Termdarstellung**[24]) wird das Konzept einer dynamischen Termdarstellung vorgestellt und seine technische Umsetzung skizziert.

Das vierte Kapitel **Termextraktion durch Beispielterme** stellt eine automatische Termextraktionsmethode vor, die mit relativ geringen Voraussetzungen gute Ergebnisse liefert und damit für weniger stark verbreitete Sprachen eine Alternative zu kommerziellen Programmen darstellt. Dieses Instrument kann bei der zentralen Aufgabenstellung der Terminografie, dem Auffinden und der Auswahl der Termini, eingesetzt werden. Hier wird aber auch gezeigt, wie die Termextraktion zur Indizierung des in den meisten terminografischen Projekten vorhandenen Hintergrundkorpus verwendet werden kann.

Das fünfte Kapitel (**Organisation rechtlichen Wissens**) gibt einen Überblick über die vielfältigen Möglichkeiten der Einteilung und Repräsentation von (rechtlichem) Wissen. Eine Methode der Wissensrepräsentation mit formaler Sprache, die logische Operationen ermöglicht, ist eine Ontologie. Es wurde eine Ontologie für eine Rechtsdatenbank erstellt und alle damit zusammenhängenden Aspekte diskutiert.

Im Fazit wird schließlich diskutiert, für welche Bereiche der Arbeit mit Rechtsdatenbanken bereits jetzt relativ einfache Werkzeuge zur Verfügung stehen und wo die Entwicklung von weiteren Werkzeugen ansetzen könnte.

Die Kapitel sind so geschrieben, dass sie auch einzeln gelesen werden können, ohne jedoch allzu starke Überschneidungen zuzulassen. Die Reihenfolge der fünf Kapitel spiegelt die Abfolge wider, in der sich die jeweiligen Aufgaben für gewöhnlich stellen: Wer Schrift im Recht mit dem Computer untersuchen will, benötigt zunächst elektronischen Text, der effizient erworben und geordnet werden sollte. Bei der Untersuchung dieser Texte und insbesondere bei der Terminografie stellt sich mit Nachdruck die Aufgabe, nicht nur ganze Gesetzesbücher, sondern einzelne Texte automatisch nach Fachgebieten ordnen zu können. Könnte der Computer diese Aufgabe nicht lösen, dann stünde jedes Ergebnis, angefangen bei der Suche nach Kontextstellen, unter einem gewichtigen Vorbehalt. Erst nachdem diese strategische Fähigkeit exemplarisch nachgewiesen worden ist, steht der Weg offen für eine vollautomatische Anpassung der Darstellung von Rechtsterminologie. Im vierten Kapitel wird dann ein Verfahren vorgestellt, das an Kapitel 1 anschließt, soweit es durch einen automatisch erstellten Index zur Korpusordnung beitragen kann, und an Kapitel 3, insoweit es die rechts-

[23] Ein Fachgebiet ist der semantische Bezugsrahmen, in dem Fachbegriffe ihre spezielle Bedeutung erhalten. Die Einordnung in Fachgebiete erfolgt zweckorientiert in Anlehnung an Konventionen. Mehr zu Fachgebieten bei Fußn. 109.

[24] ‚Term' wird in dieser Arbeit gleichbedeutend mit dem englischen Ausdruck *term* nach DIN 2342 verwendet. Damit kann sowohl nur die Benennung (aus einem Wort oder mehreren Wörtern bestehende Bezeichnung) gemeint sein, als auch der gesamte ‚Terminus' (zusammengehöriges Paar aus einem Begriff und seiner Benennung als Element einer Terminologie). Die Mehrdeutigkeit ließ sich kaum vermeiden, da sie dem Sprachgebrauch in der Computerlinguistik entspricht, die zwar eine ‚Termextraktion' oder ‚Termerkennung' kennt, aber keine ‚Benennungsextraktion' oder eine ‚Terminuserkennung'.

terminologische Arbeitsplattform erweitert. Den Schlusspunkt setzt ein Ausblick darauf, wie sich das rechtliche Wissen von der Schrift emanzipieren könnte.

Diese Arbeit wendet erstmals gewisse computerlinguistische Methoden systematisch auf elektronische Rechtsdatenbanken an. Alle Werkzeuge sind frei erhältlich und wurden erprobt und evaluiert. Neben den Resultaten werden auch die angetroffenen Schwierigkeiten und Lösungsmöglichkeiten aufgezeigt, damit diese Arbeit zu einem Leitfaden für Praktiker auf dem Gebiet werden kann.

# Literaturangaben zur Einleitung

Bungarten, T. und Engberg, J. (Hrsg.), (2003): Recht und Sprache - eine internationale Bibliographie in juristischer und linguistischer Fachsystematik, in: Hamburger Arbeiten zur Fachsprachenforschung 05, Attikon-Verlag Tostedt 2003.

Brockmeier, J. (1999): Die Welt als Bibliothek, Colloquium vom 15.7.1999, http://www.fu-berlin.de/postmoderne-psych/berichte3/brockmeier.htm : 13.11.2003.

Busse, D. (2002): Juristische Auslegungstätigkeit in linguistischer Hinsicht, S. 136-162 in: Haß-Zumkehr U. (Hrsg.) (2002), Sprache und Recht, Jahrbuch 2001 des Instituts für deutsche Sprache, de Gruyter Verlag Berlin.

Carstensen K.-U., Ebert C., Endriss C., Jekat S., Klabunde R., Langer H. (Hrsgg.) (2004), Computerlinguistik und Sprachtechnologie – Eine Einführung, Spektrum Verlag Heidelberg 2004.

DIN 2342 (1992): Begriffe der Terminologielehre, Deutsches Institut für Normung Berlin 1992.

Glenn H. P. (2000): Legal Traditions of the World, Oxford University Press New York 2000.

Henn, H., Sitta, H., Wiegand, H.E. (Hrsg.), (1998), Reihe Germanistische Linguistik 191, Max Niemeyer Verlag Tübingen 1998.

Heutger, V. (2000): Law and Language in the European Union, in: Global Jurist Topics, Vol. 3, Issue 1, Article 3, S. 4.

Holzborn T. (2003), Die Geschichte der Gesetzespublikation, Tenea Verlag Berlin 2003.

Honsell, H. (2001): Naturrecht und Positivismus im Spiegel der Geschichte, S. 593-602 in: Festschrift Koppensteiner (2001), Orac Verlag Wien 2001.

Kienast B. (1994), Die Altorientalischen Codices zwischen Mündlichkeit und Schriftlichkeit, S. 13-26 in: Gehrke H.-J. (Hrsg.)(1994), Rechtskodifizierung und soziale Normen im interkulturellen Vergleich, Gunter Narr Verlag, Tübingen, 1994.

Koch, P., Oesterreicher, W. (1994): Schriftlichkeit und Sprache, in: Hartmut Günther., Otto Ludwig (Hrsg.) (1994), Schrift und Schriftlichkeit. Ein interdisziplinäre Handbuch internationaler Forschung, 1. Halbband, Berlin/New York 1994.

Langer, S. (2002): Grenzen der Sprachenidentifizierung, S. 99-106 in: Tagungsband KONVENS 2002, DFKI Saarbrücken 2002, http://konvens2002.dfki.de/cd/pdf/19V-langer.pdf : 27.9.2004.

McEnery T. (2003), Corpus Linguistics, S. 448-463 in: Mitkov R. (Hrsg.) (2003): The Oxford Handbook of Computational Linguistics, Oxford University Press Oxford 2003.

Patzold S. (2000), Konflikte im Kloster, Matthiessen Husum 2000.

Pegoraro L., Reposo A. (1993), Le fonti del diritto negli ordinamenti contemporanei, Monduzzi Editore, Bologna, 1993.

Sieber P. (1998), Parlando in Texten. Zur Veränderung kommunikativer Grundmuster in der Schriftlichkeit, S. 182-188 in: Henn H., Sitta H., Wiegand H.E. (Hrsg.) (1998), Reihe Germanistische Linguistik 191, Max Niemeyer Verlag Tübingen 1998.

Internetquellen in Zitierreihenfolge:

http://www.univie.ac.at/RI/IRIS2004 : 30.9.2004.

http://rechtsinformatik.jura.uni-sb.de : 30.9.2004.

Kapitel 1

# I. Dokumentklassifikation beim Korpusaufbau

## Ergebnisse einer automatischen Mehrfachklassifikation beim automatischen Aufbau eines Rechtskorpus

### 1. Dokumentklassifikation als Voraussetzung für korpuslinguistische Methoden

Die Fortschritte in der Computerlinguistik und insbesondere der Korpuslinguistik eröffnen der Terminologieforschung ungeahnte neue Möglichkeiten. Terminografen werden bei der terminologischen Arbeit im engeren Sinne unterstützt, etwa bei der Zuordnung des Fachgebiets (siehe Kapitel 2). Terminologie kann u.a. mit Hilfe von Korpora effizienter erstellt und dargestellt werden (Kapitel 3). Das automatisierte Auffinden von Termkandidaten benötigt einen vorhandenen Textschatz, kann aber auch zu seiner effektiveren Verwaltung beitragen, wenn die extrahierten Terme zur Indizierung verwendet werden (Kapitel 4). Auch in vielen anderen Bereichen verspricht die Korpuslinguistik neue Anstöße für die Terminologieforschung, etwa bei der Übersetzung, dem Auffinden geeigneter Kontexte, dem Beschreiben weiterer Eigenschaften der Terme und der Verknüpfung von Terminologie.

Voraussetzung für korpuslinguistische Methoden ist stets die Erstellung eines gewichteten, annotierten und klassifizierten Korpus. Diese Aktivität erfordert Wissen und Aufwand, die viele terminologische Projekte auf den ersten Blick zu überfordern scheint. Daher wird hier als erstes eine Methode vorgestellt, mit der rechtsterminologisch relevante Dokumente semiautomatisch[25] erlangt und klassifiziert werden können. Der semiautomatische Ansatz stellt keine großen Anforderungen an den Entwicklungsgrad der Terminologie, so dass er auch für Projekte in Frage kommt, die erst in Planung oder im Aufbau sind. Darüber hinaus kann der Aufwand den Möglichkeiten des einzelnen Projekts angepasst werden, so dass sich kleine Projekte stärker auf die Automatisierung und größere Projekte mehr auf die Präzision der Methode stützen können.

Zunächst wird kurz der Forschungshintergrund für die ersten vier Kapitel vorgestellt (Punkt 2) und die Bedeutung von Korpora in der Rechtsterminologie (Punkt 3) erläutert. In Punkt 4 werden konkrete Optionen beim Aufbau eines Rechtskorpus erwogen und insbesondere grundlegende Dimensionen und Klassen zur Einteilung rechtlicher Dokumente vorgestellt (zur Einteilung in Fachgebiete siehe in Kapitel 2, zur Einteilung von rechtlichem Wissen in Kapitel 5). Die Funktionsweise der automatischen Klassifikation, wie sie auch bei der Fachgebietserkennung von Kapitel 2 eingesetzt wird, findet sich in Punkt 5. Daran schließen sich die Experimente an (Punkt 6).[26] Auf Erwägungen zur Verbesserung der Ergebnisse (Punkt 7) folgt die Zusammenfassung (8.).

---

[25] Intellektuell bedeutet, dass natürliche Personen arbeiten. Wenn sie sich dabei des Computers bedienen, spricht man von computergestützt oder semiautomatisch und wenn der Computer alleine arbeitet von automatisch oder maschinell. Luckhardt H.-D., Harms I. (2004), Virtuelles Handbuch Informationswissenschaft, Universität des Saarlandes, http://www.is.uni-sb.de/studium/handbuch//exkurs.ind.php#intellind : 19.9.2004.

[26] Die Experimente sind auf englisch beschrieben in Streiter O., Voltmer L. (2003), Text Classification for Corpus-Based Legal Terminology, S. 253-260 in: Vrabie, G., Turi, J.G. (Hrsgg.) (2003), La théorie et la pratique des politiques linguistiques dans le monde, Tagungsband der 8. Internationalen Konferenz der Académie Internationale de Droit Linguistique 2002, Editura CUGETAREA Iaşi (Rumänien) 2003.

## 2. Forschungshintergrund

In Südtirol dürfen vor Gericht und Behörden als Sprachen sowohl das Italienische als auch das Deutsche verwendet werden.[27] Um die Eindeutigkeit der Rechtssprache auch in Übersetzungen zu gewährleisten, legt man die italienisch-deutschen Übersetzungsbeziehungen per Gesetz fest.[28] Die wissenschaftliche Vorarbeit dafür leistet die Europäische Akademie Bozen (EURAC), wo seit zehn Jahren die deutsche Rechts- und Verwaltungsterminologie der italienischen Rechtsordnung aufgezeichnet und weiterentwickelt wird. Zweisprachige Terminologen, Terminografen, Fachübersetzer und Juristen aus dem italienischen, österreichischen, deutschen und schweizerischen Rechtssysteme haben im Bozner Informationssystem für Rechtsterminologie (BISTRO)[29] 15.000 Rechtskonzepte der Rechts- und Verwaltungsterminologie auf Italienisch und Deutsch beschrieben. Sie werden für Fachübersetzer und -dolmetscher, Juristen in öffentlichen Ämtern und der Anwaltschaft sowie alle an Rechtsvergleichung Interessierte im Internet veröffentlicht.[30]

Auch die ladinischsprachigen Bürger der Provinz Bozen haben das Recht, im mündlichen und schriftlichen Verkehr mit den Ämtern der öffentlichen Verwaltung in den ladinischen Ortschaften ihre Sprache zu verwenden. [31] Die Datenbank BISTRO enthält daher, in geringerem Maße, auch die ladinischen Benennungen und Kontexte für das Grödner-, Gader- und Fassatal.

## 3. Korpora in der Rechtsterminologie

Im letzten Jahrzehnt sind große elektronische Korpora allgemein verfügbar geworden und die Computerlinguistik hat sich mit großem Erfolg korpuslinguistischen Ansätzen zugewandt. Anwendungen sind *part-of-speech tagging*[32], *parsing*[33], maschinelle Übersetzung[34] und Disambiguierung[35]. Auch die Translatologie hat sehr von Translation Memories[36], Termdatenbanken und Term-

---

[27] Durchführungsbestimmungen zum Sonderstatut für die Region Trentino-Südtirol über den Gebrauch der deutschen und der ladinischen Sprache im Verkehr der Bürger mit der öffentlichen Verwaltung und in den Gerichtsverfahren, Dekret Nr. 574 des Präsidenten der Republik (DPR) v. 15.7.1988, § 48, Kapitel I, Artikel 1.

[28] Dekret des Präsidenten der (italienischen) Republik (DPR) Nr. 574 v. 15.7.1988, § 48, Kapitel II, Artikel 6, Absatz 1: „Eine mit Dekret des Regierungskommissärs gebildete paritätische Kommission aus sechs Sachverständigen, von denen drei italienischsprachige vom Regierungskommissär und drei deutschsprachige vom Landesausschuss namhaft gemacht werden:
a) bestimmt, hält auf dem neuesten Stand oder bestätigt die Rechts-, Verwaltungs- und sonstige Fachterminologie, […] um ihre Übereinstimmung in italienischer und in deutscher Sprache zu gewährleisten,
b) verfasst ein Wörterbuch der Rechts-, Verwaltungs- und sonstigen Fachterminologie in beiden Sprachen und hält es auf dem neuesten Stand."

[29] http://www.eurac.edu/bistro : 30.9.2004. BISTRO wurde präsentiert in Streiter O., Voltmer L. (2003), A Model for Dynamic Term Presentation, S. 201-204 in: Tagungsband der TIA-2003 Konferenz, LIIA – ENSAIS, Université Marc Bloch (Hrsgg.) Strasbourg 2003, http://dev.eurac.edu:8080/autoren/publs/ModDynTerm.pdf : 17.3.2004.

[30] Zur genaueren Beschreibung der komplexen Methode und ihrer praktischen Schwierigkeiten vergl. Mayer F. (2000), Terminographie im Recht: Probleme und Grenzen der Bozner Methode, S. 295-306 in: Veronesi, D., Rechtslinguistik des Deutschen und Italienischen, Unipress Padova 2000.

[31] DPR Nr. 574 v. 15.7.1988, § 48, Kapitel VI, Art. 32.

[32] Z.B. Brill, E. (1995), Transformation-based error-driven learning and natural language processing: A case study in part of speech tagging, Computational Linguistics 1995.

[33] Z.B. Charniak, E. (1996), Tree-bank grammars, S. 1031-1036 in: Proceedings of the 13th National Conference on Artificial Intelligence (AAAI-96), American Association for Artificial Intelligence (AAAI) Portland 1997.

[34] Z.B. Furuse, O. and Iida, H. (1992), An example-based method for transfer-driven Machine Translation, in: The Third International Conference on Theoretical and Methodological Issues (TMI), Empiristic vs. Rationalist Methods in MT, Canadian Workplace Automation Research Center Montréal 1992.

[35] Disambiguierung, Desambiguierung, Entambiguisierung oder Monosemierung bezeichnen die Erlangung sprachlicher oder außersprachlicher Kontextinformation zur Reduzierung lexikalischer oder struktureller Mehrdeutigkeit eines sprachlichen Ausdrucks. Ein Beispiel für den Vergleich mehrerer Methoden zur Disambiguierung beschreiben Wid-

extraktionswerkzeugen[37] sowie von den weit verbreiteten Terminologieverwaltungsprogrammen[38] profitiert.

In der Terminologie sind Korpora mehr als bloßes Ausgangsmaterial für eine Termextraktion oder einen Kontextnachweis:

- Korpora sind wichtigstes Mittel zum Verständnis eines Fachgebiets, sowohl in sprachlicher Hinsicht für den Fachterminologen als auch in inhaltlicher Hinsicht für die Fachleute.
- Korpora sind authentische Äußerungen und enthalten daher Informationen, die der Zielgruppe von Terminologie nützlich sein können.
- Korpora enthalten implizit Informationen über die Begriffe und ihre Verwendung wie idiomatische Redewendungen, Jargon, Wortwertigkeiten und Kollokationen, die meist vollständiger und aktueller sind als das in Wörterbüchern oder Termdatenbanken der Fall sein kann.
- Korpora zeigen Bedeutungsänderungen in verschiedenen Kontexten auf und sind damit ein erster Schritt in Richtung einer Begründung oder Theoriebildung der Bedeutungsverschiedenheit.
- Korpora decken auch Benennungsvarianten und Synonyme auf.

Als weiterer Punkt kommt hinzu, dass die Benutzer eine Termdatenbank oft unterschiedlich nutzen. Ein deutschsprachiger Jurist sucht andere Informationen als ein italienischsprachiger Übersetzer, so dass auch andere Darstellungsweisen als wünschenswert erscheinen. Schnell stößt die bloß veränderte Anzeige von terminologisch aufgearbeitetem Material aber auf ihre Grenzen, da eben diese Vorauswahl durch den Terminologen auf bestimmten Annahmen über die Zielgruppe beruht.[39] Die Anpassungsfähigkeit an die verschiedenen Benutzer, ein Qualitätsmerkmal im Fernunterricht und Multimediaanwendungen mit unterschiedlichen Benutzergruppen, wird heute im Wesentlichen durch geschickten Einsatz von annotierten und klassifizierten Korpora erreicht.[40]

---

dows D., Peters S., Cederberg S., Chan C.-K., Steffen D., Buitelaar P. (2003), Unsupervised Monolingual and Bilingual Word-Sense - Disambiguation of Medical Documents using UMLS, S. 9-16 in: Natural Language Processing in Biomedicine, ACL 2003 workshop proceedings, Association for Computational Linguistics Sapporo 2003, http://citeseer.ist.psu.edu/584877.html : 18.6.2004.

[36] Carl, M., Schaible, J., Pease, C. (1998), Enhancing translation memory (TM) technologies with linguistic intelligence, MULTIDOC Deliverable D 4.1 WP 6, Kommission der Europäischen Gemeinschaften Luxemburg 1998.

[37] Z.B. Bonnet, E., Gaussier, E., Langé, J.-M. (1994), A method for automatic extraction of terms from bilingual corpora, in: Proceedings of the 14th International Conference on Artificial Intelligence KBS, Expert Systems and Natural Language (AVIGNON-94), EC2 Nanterre 1994.

[38] Schmidt-Wigger, A. (1998), Building consistent terminologies, Poster in: Proceedings of the First Workshop on Computational Terminology (COMPUTERM'98) at the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics (COLING-ACL'98), ACL, Université de Montréal (Hrsgg.), Montreal 1998,
Schmidt-Wigger, A. (1998), Building consistent terminologies, in: Proceedings of the First Workshop on Computational Terminology (COMPUTERM'98) at the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics (COLING-ACL'98), ACL, University of Montreal (Hrsgg.) Montreal 1998,
http://www.iai.uni-sb.de/docs/term2.pdf : 1.10.2004. 1.10.2004.

[39] Faber, P., Lopés Rodriguez, C. I., Tercedor Sánchez, M. I. (2002), Utilización de técnicas de corpus en la representatición del conocimiento médico, S. 167-197 in: Terminology, 2002, 7(2).

[40] De Carolis, B., De Rosis, F., Pizzutilo, S. (1997), Generating user-adapted hypermedia from discourse plans, in: Lenzerini, M. (Hrsg.) (1997), Proceedings of the 5th Congress of the Italian Association for Artificial Intelligence (AI*IA97), Lecture Notes in Artificial Intelligence (LNAI), Teilreihe von: Lecture Notes in Computer Science (LNCS), 1321, Springer Heidelberg 1997.

Je größer Korpora sind, umso mehr Informationen enthalten sie.[41] Das größte Korpus ist das Internet. Das gilt mittlerweile wohl auch für rechtliche Informationen, soweit man rechtshistorische Untersuchungen ausklammert, weil die allgemeine Zugänglichkeit von Normen eine Voraussetzung für gelungene Verhaltenssteuerung ist. Normgeber versuchen daher, ihre Normen über den kostengünstigen Weg offizieller Informationsdienste im Internet zu verbreiten. Auch juristische Verlage vertreten ihre Produkte längst in elektronischer Form und geben ihre Daten zu Forschungszwecken frei, wenn sie dafür als Sponsoren aufscheinen. Die Voraussetzungen für korpusgestützte Terminografie, nämlich große, frei erhältliche Datenmengen, sind damit grundsätzlich realisierbar.

Die elektronischen Daten können entweder in ein lokales Korpus eingebracht werden oder an ihrem ursprünglichen Ort belassen und durch spezielle Indizierung und Metasuchwerkzeuge angesteuert werden. Beide Vorgehensweisen haben Vor- und Nachteile. Das lokale Korpus ist besser zu kontrollieren, d.h. es kann genauer annotiert, klassifiziert und allgemein aufbereitet werden, was auch Voraussetzung für die Ausgewogenheit eines Korpus ist. Das externe Korpus kann hingegen sehr viel größer sein und trotzdem, ohne eigenen Aufwand, stets aktuell gehalten werden.

An der EURAC wurden sowohl ein lokales wie ein externes Korpus zur Terminografie und Termdarstellung aufgebaut. Damit die Qualität der Korpora und der von ihnen getragenen terminografischen Ergebnisse gewährleistet bleibt, muss schon beim Korpusaufbau seine weitere Pflege mitgeplant werden.

## 4. Aufbau eines Rechtskorpus

### 4.1. Voraussetzung für den Einsatz von Rechtskorpora

Jedes Korpusprojekt muss die Bedingungen der Informatik, der Linguistik und des jeweiligen Fachgebiets beachten.

a) Die Bedingungen der Informatik sind:

- Wiederverwendung früherer Arbeit am Korpus,
- Sammlung der elektronischen Daten (Anfrage bei Verlagen, Suche in allgemein zugänglichen Quellen, Klärung des technischen Ablaufs),
- Datenspeicherung (technisch, aber auch verwaltungstechnisch mit verschiedenen Zugriffsrechten für Terminografen und Benutzer, für automatisierte oder manuelle Bearbeitung),
- Sammeln und Speichern von weiteren Informationen über die gesammelten Daten (Untersuchung des Dokumentinhalts durch Menschen oder computerlinguistische Methoden, Untersuchung des Dokuments selbst oder von anderen Informationsquellen über das Dokument),
- Klassifizierung der gespeicherten Daten (explizit oder implizit, physisch oder nicht, nach situationsgerechten oder allgemein anerkannten Klassifikationen),
- Aufbereitung der Daten für verschiedene Zwecke einschließlich dem Datenaustausch und der Wiederverwendbarkeit in anderen Projekten (durch Verwendung von technischen Standards und Offenheit für zukünftige Standards, indem aktuell nicht verwendete Daten, insbesondere Metadaten, aufbewahrt werden und alle Veränderungen protokolliert werden),
- Aktualisierung des Korpus (inhaltlich und technisch).

[41] Banko M., Brill E. (2001), Scaling to Very Very Large Corpora for Natural Language Disambiguation, S. 26-33 in: Meeting of the Association for Computational Linguistics 2001, http://citeseer.nj.nec.com/banko01scaling.html : 18.6.2004. In diesem Artikel wird vertreten, dass das Anhäufen von Korpora daher Priorität habe vor der Feinauswahl zwischen verschiedenen funktionierenden Methoden.

b) Die linguistische Bedingung ist:

- Ausgewogenheit des Korpus in jeglicher Hinsicht, also bezüglich aller Kategorien, nach denen die Dokumente unterteilt werden (Sprache, Fachgebiet, Rechtssystem, usw.).

c) Die fachspezifischen Bedingungen für Recht sind:

- Trennung der Rechtssprache vom Sprechen über Recht, von anderen Fachgebieten und von der Allgemeinsprache,
- Trennung von Dokumenten mit Normen, autoritativer Normauslegung (Rechtsprechung) und der Rechtslehre,
- Trennung gültiger Normen von nicht mehr gültigen und Aufzeichnung der Änderungen und ihrer Abfolge (entsprechend für Rechtsprechung und Rechtslehre, die sich auf ungültig gewordene Normen beziehen),
- Trennung von Dokumenten nach der Normenhierarchie (Verfassung, Gesetze, Satzungen usw.),
- Trennung nach dem Rang der Normgeber (EU, Staat, Land, Gemeinde usw.),
- Vereinheitlichung der Klassifikationen für die verschiedenen Rechtssysteme (Land und Regierungsbezirk in Deutschland, Region und Provinz in Italien usw.),
- Verarbeitung von ausdrücklichen und impliziten Verweisen in Dokumenten.

Es ist schwierig, allen Anforderungen gerecht zu werden, so dass verschiedene Korpusprojekte (siehe Fußnoten 70 ff.) recht unterschiedliche Lösungen gefunden haben.

4.2. Rechtliche Vorfragen eines Korpus

Man kann zwei Arten elektronischer Quellen unterscheiden. Auf der einen Seite stehen elektronische Dokumente, die uns von offizieller Seite oder von Verlagen zu Forschungszwecken zur Verfügung gestellt werden. Hat man erst einmal eine Erlaubnis zur Benutzung dieser Dokumente, dann haben sie den Vorteil, dass sie bereits bearbeitet sind und daher in gleicher, für gewöhnlich hervorragender Qualität sind. Sofern nicht bereits eine übertragbare Klassifikation vorliegt, kann fast immer die Klassifikation der gedruckten Ausgaben verwenden. Das Datenformat ist ein überwindbares Problem, im Gegensatz zum Kopierschutz. Die Daten können bei einem benutzerorientierten und interaktiven Ansatz nicht intern bleiben, sondern müssen auch den Benutzern zugänglich gemacht werden. Genau dies macht aber die Arbeit, mit der Verlage ihr Geld verdienen, allgemein zugänglich. Selbst Gesetzestexte, die als solche nicht geschützt sind, fallen durch ihre Aufbereitung (Unterteilung, Verknüpfung usw.) unter Urheberschutz. Dokumente von Verwaltungen können datenschutzrechtlich geschützt sein. Selbst wenn man eine Zustimmung erhalten würde, wäre ein so erstelltes Korpus ständig von einem Widerruf der Genehmigung bedroht und würde das gesamte Projekt auf tönerne Füße stellen.

Auf der anderen Seite gibt es öffentlich zugängliche elektronische Dokumente von hoher Qualität und Aktualität im Internet. Das Kopieren ganzer Datenbanken und das Darstellen fremder Inhalte auf eigenen Seiten (sog. *framing*) ist in den meisten Rechtssystemen verboten. Suchmaschinen (z.B. GOOGLE) kopieren zwar im großen Stil Internetseiten und haben bisher noch keine rechtlichen Konsequenzen zu spüren bekommen, das muss allerdings nicht auf Dauer so bleiben.[42] Texte

[42] Einen Überblick zu *framing*, *inline linking* und *deep framing* bietet Ott S., JurPC Web-Dok. 14/2003, http://www.jurpc.de/aufsatz/20030014.htm : 1.10.2004. Problematisch wird es spätestens dann, wenn Inhalte angezeigt werden, die der Autor aus dem Verkehr gezogen hat, denn damit entfällt jede Basis für die Konstruktion einer stillschweigenden Zustimmung zur Indexierung dieser Inhalte.

in den verbreiteten Sprachen kann man im Internet jedoch aus vielen verschiedenen Quellen in ausreichender Menge herunterladen.

Auf den ersten Blick scheint weder der rechtlich sichere, aber tatsächlich dornenreiche Weg über Verlage und offizielle Stellen noch der effektive, aber rechtlich verfängliche Weg durch Kopieren fremder Inhalte gangbar zu sein. Andererseits ist das bloße Verweisen auf Internetinhalte wegen der Unbeständigkeit der Seiten keine Lösung. Weil Seiten zu rasch ihre Adresse und ihren Inhalt verändern, ist ein lokal gespeichertes Korpus fast unumgänglich.

Hier hilft das wissenschaftliche oder publizistische Zitat, das mittlerweile von der Rechtssprechung als stets erlaubt anerkannt zu werden scheint.[43] Das Darstellen von Zitatbruchstücken der elektronischen Dokumente dürfte damit sowohl für die Verwendung von Verlagsmaterial wie auch von nicht mehr veröffentlichten Internetseiten rechtlich gedeckt sein. Als einzige rechtliche Gefahr bleiben mögliche Schadensersatzansprüche durch Verlage, wenn deren Material über die Korpusdarstellung von Dritten geraubt wird.[44] Damit reduziert sich die Wahl auf Internetdokumente, die allerdings nicht ganze Datenbanken umfassen dürfen und deren Inhalte nicht als eigene dargestellt werden dürfen.

4.3. Korpusaufbau

Internetdokumente können durch einen Web-Robot[45] lokal gespeichert werden. Dazu gibt man eine beliebige URL[46] ein. Der Robot holt sich die Seite von der angegebenen Adresse, prüft sie auf die angegebenen Mindestanforderungen (Kontrolle der robots.txt-Datei auf Kopierschutz, Kontrolle der *robots meta-tags* auf Erlaubnis zum Speichern der Verweise, Mindestzeichenzahl, Sprache usw.) und speichert den Inhalt der Seite dann lokal ab. Die Verweise der gespeicherten Seite werden in eine Einkaufsliste geschrieben und weiter bearbeitet. Bilddateien wie .gif, .tif, .jpg oder cgi-Programme des Servers werden an der Dateinamenerweiterung erkannt und von der Liste gestrichen. Gehen die Verweise auf geeignete Seiten, werden auch sie gespeichert und ihre Verweise zur Einkaufsliste hinzunotiert. Dieses Schneeballsystem kann und muss auf geeignete Weise gesteuert werden. Dazu kann die Suche auf eine Domain und auf Schlagwörter beschränkt werden und eine maximal zu speichernde Seitenzahl angegeben werden.[47] Außerdem wird ein zeitlicher Mindestabstand für die automatischen Zugriffe eingebaut, um die Server nicht zu überlasten. Der nötige Web-Roboter kann im Internet frei heruntergeladen werden.[48]

Dieser Web-Roboter eignet sich auch vorzüglich zum Aktualisieren des Korpus. Es genügt, die Anfragen zur Erstellung zu speichern und nach einiger Zeit erneut in dieser oder ähnlicher Weise selbständig laufen zu lassen. Der Robot ist in der Lage zu erkennen, ob ein Dokument bereits gespeichert ist und ob der Inhalt noch derselbe ist. Dann wird es nicht erneut gespeichert.

---

[43] In diese Richtung geht der BGH für deutsches Recht im Urteil v. 11.07.2002, I ZR 255/00, „Elektronischer Pressespiegel“ JurPC Web-Dok. 302/2002, http://www.jurpc.de/rechtspr/20020302.htm : 1.10.2004, wo allerdings eine Volltextrecherche zum Schutz der Informationshersteller ausgeschlossen sein muss.

[44] Der Verlag könnte argumentieren, dass die überlassenen Inhalte nicht ausreichend gegen solche Angriffe gesichert wurden. Nach deutschem Recht könnte das eine positive Vertragsverletzung (pVV) des Überlassungsvertrags darstellen.

[45] Auch bot, *crawler*, Agent oder Spider genannt. Programm, das einen Auftrag nach dem Initialisieren selbständig weiter ausführt.

[46] URL steht für *unified resource locator* und ist die Adresse eines Dokuments im Internet. Die Adresse besteht regelmäßig aus dem Namen des Protokolls (z.B. http, ftp, dap, file), der Domain, dem Verzeichnis und dem Dateinamen.

[47] Weitere Einschränkungen können sein, ob die an der URL ablesbare Hierarchie nach oben weiterverfolgt werden darf und ob eine bestimmte Zeichenfolge wie „Recht“ in der URL vorkommen soll. Die hierarchische Strukturierung von Domains hilft oft, nur auf die gewünschten Dokumente zuzugreifen. Wenn Textdateien nur am Ende der Hierarchie zu finden sind, wird die Hierarchie bis nach unten abgefragt und wenn die Dokumente alle auf paralleler Hierarchieebene gespeichert sind, kann auch dies angegeben werden.

[48] Z.B. wget von Open Source Robot: http://www.gnu.org/manual/wget/: 23.9.2003.

Zu jeder Seite werden das zum wissenschaftlichen Zitieren nötige Datum und die für die Suche verantwortliche Person als Metadaten gespeichert. Diese Metadaten werden wie unten beschrieben zur Klassifikation verwendet.

4.4. Kodierung des Korpus

In einem Korpus werden Dokumente und Informationen über die Dokumente gespeichert. Anfragen an das Korpus sollten über alle Daten gleichzeitig laufen können, was am leichtesten durch ein gemeinsames Datenformat zu gewährleisten ist. an der EURAC wurde XML gewählt, weil XML der Standard für Datenrepräsentation ist[49] und weil viele Werkzeuge für dieses Format kostenlos im Web erhältlich sind[50], so dass nicht alle korpuslinguistischen Werkzeuge selbst erstellt werden müssen.

Alle Daten müssen also in XML-Format umgewandelt werden. Relativ einfach umzuformen sind bibliographische Angaben, die im MAB2-Standard[51] von Online-Verbundkatalogen abgerufen werden.[52] Internetdokumente haben aber verschiedene Formate, z.B. .html, .pdf und .txt. Daher wurde ein Konvertierungsprogramm geschrieben, das zunächst alle Formate in HTML und dann in XML umformt.

Ein weiterer Standard, der eingehalten werden sollte, ist der CES (*corpus encoding standard*) Standard für die Kodierung von Korpora und dessen XML-Version XCES.[53] CES normiert die Angaben über die Annotierung der Dokumentstruktur (Titel, Überschrift, Absatz Satz), über linguistische Annotierung im Dokumenttext (flektierte Formen, Phrasen) und über Textalinierung[54]. Für die Speicherung dieser Daten im Header gibt es wiederum einen Standard, den TEI (*text encoding initiative*), der normiert, wie und wo die *tags* sein sollen, in denen die Angaben stehen.[55] Die Standards sind alle untereinander kompatibel, so dass die Daten allen Standards gleichzeitig entsprechen können und sollten.

Nun müssen die standardmäßig vorgesehenen Felder natürlich noch ausgefüllt werden. Internetdokumente enthalten in den seltensten Fällen überhaupt verwendbare Metadaten, geschweige denn in standardisierter Form. Außerdem müssen Metafelder nicht nur beim Aufbau eines Korpus ausgefüllt werden, sondern bei jedem neu hinzukommenden Dokument, ja, bei jeder Änderung der Inhalte der Dokumente und ihrer Beziehungen. Man sollte also computerlinguistische Hilfen so weit wie

---

[49] Extensible Markup Language, http://www.w3.org/XML: 23.9.2003.

[50] Beispielsweise http://xml.apache.org: 23.9.2003.

[51] MAB ist das Maschinelle Austauschformat für Bibliotheken. MAB ist der Standard für Darstellung und Austausch von bibliographischen Daten, Normdaten, Lokaldaten und Besitznachweisen in Deutschland. Das MAB2-Format ist das deutsche Metadatenformat für alle in den deutschsprachigen Ländern genutzten bibliographischen Datenformate. Die MAB2 Felder sind dreistellige Nummern. die offizielle Beschreibung findet sich unter http://www.ddb.de/professionell/mab.htm : 21.9.2004.

[52] Dazu wurde ein *mapping*-Programm geschrieben, das mit der ISBN-Nummer beim Verbundkatalog (http://www-opac.bib-bvb.de :21.9.2004) die Daten bestimmter Felder des MAB2-Standards abruft und automatisch in die entsprechenden XML-Angaben überführt. Das Programm läuft unter http://dev.eurac.edu:8080/cgi-bin/bib/biblio : 18.6.2004 und kann kostenlos online benutzt werden. Da die Klassifikationsdaten bereits in elektronischer Form übermittelt werden, können sie sofort zur Klassifikation verwendet werden.

[53] XCES ist der Korpuskodierungsstandard (*Corpus Encoding Standard* - CES) auf XML Basis und basiert auf Kodierungsstandards für linguistische Korpora, vergl. http://www.cs.vassar.edu/ces : 4.10.2004 und und http://www.cs.vassar.edu/XCES : 4.10.2004.

[54] Alinierung ist die Angabe, welche Textstellen einander entsprechen, insbesondere welcher Text bzw. welches Wort die Übersetzung zu einem Ausgangstext bzw. -wort ist.

[55] Ein TEI-Header muss Angaben über die elektronische Quelle, das entsprechende Druckwerk und das Verhältnis der beiden zueinander machen. Man kann Bibliographische Angaben als Druckwerke eingeben, deren elektronische Version nicht im Korpus ist. Das hat nicht nur den Vorteil, dass später nur noch die elektronische Version angegeben werden muss, sondern auch, dass Header für Dokumente von Verlagen und aus dem Internet gleich aussehen und behandelt werden können. Mehr dazu unter http://www.tei-c.org/: 4.10.2004.

möglich ausnutzen. Im Folgenden werden Methoden zur automatischen Erstellung der nötigen Metadaten für Dokumente aus dem Internet vorgestellt.

4.5. Welche Klassen für rechtliche Dokumente?

Zunächst sei darauf hingewiesen, dass es hier ganz pragmatisch um die Einteilung von rechtlichen Dokumenten geht, und nicht um die Einteilung von rechtlichem Wissen, das in diesen Dokumenten enthalten sein mag (oder nicht). Mit der Einteilung von rechtlichem Wissen, einem sehr komplexen Thema, wird sich Kapitel 5 (Wissensorganisation) ausführlich beschäftigen. Für den Korpusaufbau muss die Einteilung aber den Anforderungen einer Informationssuche durch die geeignete Repräsentation der Dokumente entsprechen[56] und für alle Nutzer unmittelbar durchschaubar sein. Da die Nutzer sowohl Rechtsexperten wie Linguisten sind, dürfen die Klassen nicht nur rechtlicher Natur sein. Ideal wären Metadaten, die einem Standard entsprechen und die zugleich als wissenschaftliche Quellenangabe[57] dienen können.

Existierende Katalogisierungsstandards[58] kommen mit einem einzigen, tief verzweigten Klassifizierungsbaum aus. Sie passen aber nicht richtig zur Aufgabenstellung, denn sie beschäftigen sich nur mit Druckwerken, klassifizieren aus der Sicht eines Außenstehenden und vor allem gehen sie von der Annahme aus, dass der zu klassifizierende Inhalt „der selben Welt“ angehört. Rechtstext bezieht sich aber nicht auf die konkrete Welt, sondern auf eine Ideenwelt und darüber hinaus auf eine bestimmte von vielen konkurrierenden Ideenwelten. Diese Ideenwelt heißt Rechtssystem. Normen aus verschiedenen Rechtssystemen gehen fast immer völlig beziehungslos aneinander vorbei. Wenn ein österreichischer Jurist nach § 12 Einkommensteuergesetz sucht, dann hilft ihm § 12 des italienischen Einkommensteuergesetzes nichts. Ein rechtsvergleichender Standard zur Klassifikation von Dokumenten aus verschiedenen Rechtsordnungen hat sich noch nicht herausgebildet und die bisherigen Ansätze in der Wissenschaft sind spärlich. Zu nennen sind immerhin die Arbeiten von Engberg[59], Frilling[60] und Lundquist[61]. In diesen Untersuchungen wird versucht, eine bestimmte Auswahl von Rechtstexten, deren Ordnung als unproblematisch vorausgesetzt wird, nach semantisch-lexikalischen Merkmalen einzuteilen. Hier ist das Ziel hingegen eine Einteilung, die allen unter Punkt 4.1. genannten Voraussetzungen für alle möglichen Rechtstexte gerecht wird. Folgende fünf Kriterien sind für eine Klassenbildung besonders interessant:

- Sprache des Dokuments,
- Rechtsordnung, auf die sich der Inhalt bezieht,
- Rechtsqualität des Dokumentinhalts, also ob es eine Norm, eine offizielle Norminterpretation (Urteile etc.), oder nichtoffizielle Informationen sind,[62]

---

[56] Vergleiche Kapitel 4 Punkte 6 und 11 zur Textindexierung.

[57] Wie oben angesprochen kann grundsätzlich jedes Dokument mit seiner URL zitiert werden. Oft soll aber unabhängig von der Veröffentlichung jener Seite auf den weiter bestehenden (Norm-)Inhalt verwiesen werden. Korrekter wäre dann die Angabe der Norm oder offiziellen Fundstelle (v.a. bei Urteilen), damit der wahre Autor an der Quellenangabe ersichtlich ist.

[58] Z.B. UDC, DDC, Dewey oder Regensburger Verbundklassifikation http://www.bibliothek.uni-regensburg.de/rvko_neu/mytree.php3#P: 27.9.2003.

[59] Engberg, J. (1993), Prinzipien einer Typologisierung juristischer Texte, S. 31-38 in: Fachsprache 15 1/2, Wien 1993 und zuletzt Rasmussen, K. W; Engberg, J. (1999), Genre Analysis of Legal Discourse, S. 113-132 in: Hermes - Journal of Linguistics 1999, 22.

[60] Frilling S. (1994), Textsorten in juristischen Fachzeitschriften, Internationale Hochschulschriften 138, Waxmannverlag Münster/New York 1995, zugl. Diss. Univ. Münster 1994.

[61] Lundquist, L. (1979), Teksttypbestemmelse af en lovtekst via en semantisk dybdestruktur, in Linnarud, M., Svartvik, J. (Hrsgg.) (1979): Kommunikativ Kompetens och fackspråk, SYMPOSIUM Södertälje 1978, ASLA Uppsala 1979.

[62] Vergleiche die ‚Dokumenttypen’ bei Unger W., „Methoden juristischer Dokumentrecherche“, http://www.juralink.de/8LITERATUR/Umgang/Recherche.htm: 27.9.2003.

- falls Norm oder Norminterpretation: die Ebene in der Normen- und Staatsorganisationshierarchie, aus der die Norm oder Norminterpretation hervorgegangen ist,
- Rechtsgültigkeit, also ob die Norm, auf die sich der Inhalt bezieht, noch Gültigkeit besitzt und
- das Fachgebiet.

Entscheidend sind aber nicht nur die Kriterien der Klassen, sondern auch die Festlegung der Kategorien: Was wird als Sprache des Dokuments eingegeben, wenn der Text mehrsprachig ist? Was, wenn eine Norm für mehrere Rechtssysteme gilt oder wenn sie nur subsidiär anwendbar ist? Was, wenn in einem nichtoffiziellen Dokument eine offizielle Norm veröffentlicht wird? Entsprechen die deutschen und österreichischen Bundesländer den schweizer Kantonen und alle ihrerseits einer italienischen Region? Was ist der Gültigkeitsstatus, wenn die Norm ab oder bis zu einem bestimmten Datum gilt?

Für Sprachen bietet es sich unmittelbar an, die geltenden ISO-Normen zu verwenden.[63] Diese kennen keine Sprachkombinationen, so dass die Kodierung eines mehrsprachigen Texts durch mehrfache Etikettierung erfolgt. Das entspricht in aller Regel auch den Bedürfnissen der Benutzer, die auf alles Sprachmaterial einer Sprache zugreifen möchten, selbst wenn es in einem mehrsprachigen Text steht,[64] vor allem bei weniger verbreiteten Sprachen wie Grödnerisch, Gadertalerisch, Fassanisch und Standard Dolomitenladinisch[65].

Am schwierigsten ist die Lösung für die Gültigkeit von Normen. Normen können formal noch gültig, aber obsolet sein. Bei Normenkontrollverfahren kann gerade die Gültigkeit einer Norm Streitgegenstand sein. Wenn selbst Fachleute nicht jederzeit eindeutig die Gültigkeit einer Norm feststellen können, wie soll es dann bei der Korpusklassifikation oder gar automatisch geschehen? Die einzige Information, die sich in aller Regel auf die Gültigkeit bezieht und die automatisch herausgefunden werden kann, ist die Chronologie von aufeinander folgenden Normen. Wenn im Korpus zwei Versionen eines Gesetzes sind, dann spricht viel dafür, dass das neuere Gesetz das alte ersetzt hat. Damit ist das alte ungültig und das neue gültig. Mehr Informationen über die Gültigkeit würden bei der korpuslinguistisch erforderlichen Menge an Dokumenten die Datenpflege überfordern.

Als Rechtsordnung wurden alle souveränen Staaten gewählt und, juristisch fehlerhaft aber aus informationstechnischen und praktischen Gründen, die beiden Auffangkategorien „EU-Recht“ und „international“. Ein Dokument kann sich zugleich auf mehrere Rechtsordnungen beziehen.[66]

Die Ebene in der Normen- und Staatsorganisationshierarchie wurde nur sehr schwach an die *territorial units for statistics* (NUTS)[67] der EU angelehnt und werden eher untechnisch und pragmatisch zur weiteren Unterteilung angewandt. Da nur offizielle Normen und Norminterpretationen hierarchisch zueinander stehen, gibt es nur einen oder keinen Eintrag in dieser Klasse.

---

[63] Die ISO 639 normiert, wie Sprachen abzukürzen sind. Es gibt Abkürzungen mit zwei Buchstaben und mit drei Buchstaben, und es gibt leicht abweichende Versionen bibliografische (ISO 639-2/B) und für terminologische Zwecke (ISO 639-2/T). http://www.loc.gov/standards/iso639-2/normtext.html : 1.10.2004.

[64] Es gibt also keine eigene Klasse „deutsch-italienisch“, sondern ein Dokument kann mehreren Klassen angehören. Diese Klassifizierung könnte verbessert werden, indem man die Klassifizierung nur auf den Teil des Textes anwendet, der tatsächlich in der jeweiligen Sprache geschrieben ist, so dass keine falschen Treffer für die weitere Sprache entstehen.

[65] Grödnerisch wird im Grödnertal gesprochen, Gadertalisch oder Gadertalerisch im Gadertal, Fassanisch im Fassatal. Das Standard Dolomitenladinisch (*Ladin dolomitan* oder *Lingaz standard di Ladins dles Dolomites*) wurde vom *Servisc de planificazion y elaborazion dl lingaz ladin* (SPELL) als gemeinsame Sprache aller fünf (nach anderer Einteilung acht) Varianten des Dolomitenladinischen aus der Taufe gehoben.

[66] Z.B. ein Rechtsvergleich oder eine internationale Norm, die zugleich unmittelbares anwendbares Recht in mehreren Staaten ist.

[67] http://europa.eu.int/comm/eurostat/ramon/nuts/codelist_en.cfm: 27.9.2003.

**Tabelle 1: Aufstellung der Klassifizierung**

| Sprache | Rechtsordnung | Normen- und Staatsorganisationshierarchie | Rechtsqualität |
|---|---|---|---|
| Italienisch | International | Staat | Gesetz |
| Deutsch | EU | selbständiger Staatsteil/Teilstaat | andere Norm |
| Grödnerisch | Italien | Region/Provinz | Urteil |
| Gadertalerisch | Österreich | Gemeinde | anderes Rechtsdokument |
| Fassanisch | Deutschland | Restkategorie | Wörterbuch |
| Standard Dolomitenladinisch | Schweiz | | Anderes Dokument |

Die weitaus komplexeste und am stärksten am Inhalt orientierte Unterteilung ist diejenige in Fachgebiete.[68] Eine erste Schwierigkeit besteht schon darin, dass verschiedene Fachgebietseinteilungen konkurrieren. Zum einen konkurriert die korpuslinguistische Einteilung (gleich große Datenmengen je Klasse) mit der sprachlichen Einteilung (eigenes Fachvokabular und eigene Ausdrucksweise) und der rechtlichen Einteilung (allgemeines Schuldrecht und besonderes Schuldrecht unterscheiden sich stark in der Datenmenge, haben geringe linguistische Unterschiede in ihrer Fachsprache, aber eine je eigene rechtssystematische Stellung). Im Kapitel 2 wird noch genauer auf die Problematik der Fachgebietseinteilung eingegangen.

Jedenfalls konnte sich weder eine Norm für die Klassifizierung von Recht, noch ein verbreiteter Gebrauch für einen bestimmten Bereich durchsetzen. Jede Klassifizierung wird daher höchst angreifbar bleiben, was für eine pragmatische und nicht allzu tief verzweigte Einteilung spricht. Von Vorteil erscheint jedenfalls eine Dezimalklassifikation, um flexibel hinsichtlich der Einteilung und Bezeichnung der Klassen zu bleiben. Da an der EURAC terminologische Einträge bereits in Fachgebiete eingeteilt waren, wurden zunächst einmal diese auch für die Einteilung von Korpusdokumenten verwendet.

Eine wichtige Entscheidung ist, ob die Fachgebiete alle nebeneinander stehen oder hierarchisch aufgebaut sind.[69] Wenn es eine Hierarchie gibt ist wichtig, ob nur die jeweils niedrigste, konkreteste Ebene Dokumente enthalten darf, so dass weiter unterteilbare Klassen kein eigenes Fachgebiet, sondern nur Unterteilungskriterien sind. Klarer ist es, wenn alle Klassen mit Dokumenten besetzt werden können und nicht zur bloßen Orientierung eingeführt werden, weil dann alle Knoten gleichwertig sind. Eine durchgehende Entscheidung ist ganz allgemein zu bevorzugen, so dass bei einer Hierarchie wirklich alle Klassen zueinander in Beziehung gesetzt werden. Für eine Vertiefung der Ordnungsproblematik wird auf Kapitel 5 verwiesen.

Die für die Versuche zur automatischen Klassifikation verwendete hierarchische Einteilung orientiert sich an der Systematik des Bundesrechts des Juristischen Internetprojekts Saarbrücken[70], an

[68] Siehe ausführlichere Diskussion bei der Fachgebietserkennung in Kapitel 2. Hier geht es um allgemein anerkannte und intuitiv für alle Nutzer erfassbare Rechtsgebiete.

[69] Zu den verschiedenen Arten von Hierarchien vergl. Kapitel 5 bei Fußn. 320 f.

[70] http://www.jura.uni-sb.de/BGBl/BGBLSYST.HTML : 1.10.2004. Herberger, M., Systematik des Bundesrechts, Projektbericht, Juristisches Internetprojekt Saarbrücken, makrolog Wiesbaden 1997 (Internetausgabe). Auszüge auf nächster Seite. Diese Klassifikation dürfte die Materialmenge in den jeweiligen Fachgebieten berücksichtigen und scheint ein Kompromiss zwischen den korpuslinguistischen und fachlichen Bedürfnissen zu sein. Eine fachliche Schwäche ist die Kategorie ‚4 Zivil- und Strafrecht', die trotz zehn Klassen disparate Rechtsgebiete vermengt und statt dessen Unterteil-

der Systematik der Schweizer Gesetzestexte Online[71] und der *Documentazione Giuridica*[72] für Italien. Die Hierarchie geht aus den Dezimalzahlen in der Tabelle 2 hervor: Untergebiet von 2 (Verwaltungsrecht) ist 21 (besonderes Verwaltungsrecht), und Untergebiete von 21 sind beispielsweise 211 (Straßenverkehrsrecht) und 213 (Baurecht).

Trotzdem kann die automatische Klassifikation die Hierarchie ignorieren und alle Klassen gleich behandeln. Für die Veranschaulichung ist die Information über die Beziehung der Klassen zueinander sicher wertvoll, sie verkompliziert aber die Automatisierung. Daher bleibt sie zunächst außer Betracht und kann bei einer späteren Verfeinerung der Methoden eingebaut werden.

lungen des Zivilrechts in den Rang von Hauptklassen (eine Dezimale) erhebt, z.B. ‚7 Wirtschaftsrecht' oder ‚8 Arbeits- und Sozialrecht'.

[71] http://www.gesetze.ch/: 27.9.2003.

[72] http://www.idg.fi.cnr.it/banche/dogi/dogi.htm: 10.10.2003.

**Tabelle 2: Systematik des Bundesrechts**

| | |
|---|---|
| 0 | Recht allgemein |
| 1 | Öffentliches Recht |
| 10 | Verfassungsrecht |
| 11 | Staatsorganisationsrecht |
| 17 | Europarecht |
| 18 | Völkerrecht |
| 2 | Verwaltungsrecht |
| 21 | besonderes Verwaltungsrecht |
| 211 | Straßenverkehrsrecht |
| 213 | Baurecht |
| 219 | Polizeirecht |
| 221 | Universitätsrecht |
| 3 | Rechtspflege |
| 30 | Gerichtsverfassungsrecht |
| 31 | Prozessrecht |
| 310 | Zivilprozessrecht |
| 312 | Strafprozessrecht |
| 315 | freiwillige Gerichtsbarkeit |
| 318 | Verwaltungsprozessrecht |
| 4 | Zivil- und Strafrecht |
| 40 | bürgerliches Recht |
| 401 | bürgerliches Recht allgemeiner Teil |
| 402 | Schuldrecht |
| 403 | Sachenrecht |
| 404 | Familienrecht |
| 405 | Erbrecht |
| 41 | Handelsrecht |
| 411 | Börsenrecht |
| 412 | Gesellschaftsrecht |
| 45 | Strafrecht |
| 5 | Militärrecht |
| 6 | Finanzwesen |
| 7 | Wirtschaftsrecht |
| 76 | Geld-, Kredit- und Versicherungswesen |
| 79 | Umweltrecht |
| 8 | Arbeits- und Sozialrecht |
| 80 | Arbeitsrecht |
| 86 | Sozialrecht |
| x | nicht Fachgebiet Recht |

## 5. Wie funktionieren automatische Klassifizierer?

### 5.1. Ausgangsdaten für automatische Klassifizierer

Automatische Dokumentklassifikationen nutzen in aller Regel die Ähnlichkeit des zu klassifizierenden Dokumentes mit bereits klassifizierten Dokumenten.[73] Beim Aufbau eines Korpus sind al-

[73] Chien, L.-F., Chen, C.-L. (2001), Incremental extraction of domain-specific terms from online text resources, S. 89-109 in: Bourigault, D., Jacquemin, C., L'Homme, M.-C. (Hrsgg.) (2001), Recent Advances in Computational Terminology (Natural Language Processing), John Benjamins Amsterdam 2001.
Nohr, H. (2001), Automatische Indexierung, Einführung in betriebliche Verfahren, System und Anwendungen, Materialien zur Information und Dokumentation, Verlag für Berlin-Brandenburg Potsdam 2001.

lerdings noch keine klassifizierten Dokumente vorhanden, die man zum Training automatischer Verfahren verwenden könnte. Damit scheint ein automatischer Ansatz von vornherein ausgeschlossen zu sein. Es gibt jedoch bestimmte Informationen, die beim Datensammeln selbst anfallen, die von automatischen Verfahren genutzt werden können. Solche Daten sind in Tabelle 3 verzeichnet:

**Tabelle 3: Information und Herkunft**

| Information | Art und Herkunft |
|---|---|
| URL des Dokuments | Adresse des Dokuments aus dem Link |
| Schlagwörter | in den Metadaten eines Dokuments |
| Inhaltsbeschreibung | in den Metadaten eines Dokuments |
| Titel | im Dokumentinhalt und den Metadaten |
| Verweisadresse im Dokumentinhalt | URL, auf die ein Link verweist. |
| Mitarbeiterprofil | Mitarbeiter-Loginname, Startzeit des Web-Robots |

Diese Daten müssen vom Web-Robot jeweils getrennt gespeichert werden. Zum Ähnlichkeitsvergleich bietet es sich ganz offensichtlich an, die URLs zweier Dokumente heranzuziehen. Auch ein Vergleich ungleicher Klassen wie etwa Schlagwörter und Inhaltsbeschreibung könnte aber Sinn machen. Chakrabarti et al. zeigen beispielsweise, dass der Vergleich der URL eines Dokuments mit der Verweisadresse[74] eines anderen ein wertvolle Informationen über deren Ähnlichkeit enthält.[75] Es ist auch nachvollziehbar, dass eine Seite auf eine sehr ähnliche Seite verweist, die ihrerseits nicht mehr auf ähnliche Seiten weiter verweist, so dass ein solcher Klassifikator zusätzliche Informationen einbringen kann.

Hier werden also auch die URL des zu klassifizierenden Dokuments mit den Verweisadressen in den klassifizierten Dokumenten und die Verweisadressen des zu klassifizierenden Dokuments mit den URLs der klassifizierten Dokumente verglichen.

[74] Mit dem Open Source Browser LYNX kann man die Verweisadressen leicht vom Dokumentinhalt trennen: http://lynx.browser.org: 10.10.2003. Die Verweisadresse in <a href= “http://www.eurac.edu“> EURAC </a> ist http://www.eurac.edu. Man könnte auch den Verweistext (Im Beispiel wäre das „EURAC“) mit einbeziehen, wie das die Suchmaschine GOOGLE tut, um Suchergebnisse zu sortieren: „link structure and link text provide a lot of information for making relevance judgments and quality filtering“ Brin, S., Page, L. (1998), The anatomy of a large-scale hypertextual Web search engine, S. 107-117 in: Computer Networks and ISDN Systems Vol. 30, Nr.1-7 1998, http://citeseer.nj.nec.com/brin98anatomy.html : 1.10.2004. Mit Hilfe von GOOGLE könnte man sogar die in-links, also die Seiten, die auf eine URL hinverweisen, bekommen. Hier wurde auf die weitere Abfrage von Daten verzichtet, es handelt sich also immer um out-links, die wegverweisen.

[75] Chakrabarti, S., Dom, B., Indyk, P. (1998), Enhanced hypertext categorization using hyperlinks, S. 307-318 in: Haas L. M., Tiwary A. (Hrsgg.) (1998), Proceedings ACM SIGMOD International Conference on Management of Data, ACM Press Seattle 1998, untersuchten allerdings die Identität von Verweisadresse und URL. Es scheint offensichtlich, dass die konkret gemeinte Seite inhaltlich mit der Ausgangsseite zu tun hat. Sehr viel gewagter ist die Hypothese, dass Seiten, die ähnliche *n*-Gramme in der Adresse haben, dem verweisenden Dokument statistisch gesehen ähnlicher sind als eine Zufallsauswahl.

**Tabelle 4: Informationsvergleiche zur Vorhersage**

| Vergleich | erhoffte Information |
|---|---|
| Dokumentinhalt – Dokumentinhalt | Sprache, Rechtsordnung, Fachgebiet, interne Hierarchie |
| URL - URL | Sprache, Rechtsordnung, Fachgebiet, interne Hierarchie |
| Schlagwörter - Schlagwörter | Sprache, Fachgebiet, interne Hierarchie |
| Inhaltsbeschreibung - Inhaltsbeschreibung | Sprache, Fachgebiet, interne Hierarchie |
| Titel - Titel | Sprache, Rechtsordnung, Fachgebiet, interne Hierarchie |
| Verweisadresse - Verweisadresse | Sprache, Rechtsordnung, Fachgebiet |
| Mitarbeiterprofil - Mitarbeiterprofil | Fachgebiet[76] |
| URL – Verweisadresse[77] | Sprache, Fachgebiet |
| Verweisadresse - URL[78] | Sprache, Fachgebiet |

5.2. Kombination von Klassifikatoren

Die neun vorgestellten Klassifikatoren (siehe Tabelle 4) werfen nicht nur die Frage nach dem besten Klassifikator auf, sondern auch danach, ob eine Kombination der Klassifikatoren bessere Ergebnisse verspricht als jeder der Klassifikatoren für sich alleine.[79] Für die Kombination von eigenständigen Klassifikatoren ist seit langem anerkannt, dass selbst gute Klassifikatoren durch eine Kombination schwacher Klassifikatoren geschlagen werden können.[80] Nach jüngeren Forschungsergebnissen scheinen kombinierende Systeme ganz allgemein einzelnen Klassifikatoren überlegen zu sein.[81] Es gibt zwar viele Kombinationsmöglichkeiten: „The designer can create a myriad of different MCSs [multiple classifier systems] by coupling different techniques to create classifier ensembles with different combination functions“, aber noch keine allgemein akzeptierte Strategie für die Kombination.[82] Beispielsweise führt das Kombinieren von mehreren Ähnlichkeitskriterien in einem einzigen Klassifikator[83] zum *peaking*-Phänomen: Mit den Ähnlichkeitskriterien müssen die Trainingsdaten exponentiell zunehmen, damit die Ergebnisse nicht schlechter werden.[84] Um diesen

[76] Die einzelnen Mitarbeiter sind mit einem je eigenen Fachgebiet betraut und werden Dokumente für ihre eigene Arbeit speichern.

[77] Hinverweisender Link oder *in-neighbour*.

[78] Wegverweisender Link oder *out-neighbour*.

[79] Die Kombination hat sich etwa bei der *optical character recognition* bewährt: Schulz, K. U. (2003), Nachkorrektur von Ergebnissen einer optischen Charaktererkennung, S. 29-38, http://www.cis.uni-muenchen.de/people/Schulz/SkriptOCR_03/OCR.pdf: 4.4.2005.

[80] Larkey, L. S., Croft, W. B. (1996), Combining classifiers in text categorization, S. 289-297 in: Proceedings of SIGIR-96, 19. Internationale Konferenz über Wissenschaft und Entwicklung im Information Retrieval, ACM Press, New York/Zürich 1996. In diesen Versuchen ist die Kombination beliebiger Klassifikatoren besser als der beste Klassifikator alleine. Die Klassifikatoren müssen dafür wohl aber hinreichend unterschiedlich sein, damit ihre jeweilige Stärke die Schwäche des Partnerklassifikators wettmachen kann.

[81] Kang H.J., Doermann D. (2003), Evaluation of the Information-Theoretic Construction of Multiple Classifier Systems, Proceedings of the 7th International Conference on Document Analysis and Recognition (ICDAR 2003), IEEE Edinburgh 2003, http://citeseer.ist.psu.edu/ps/698712 : 26.9.2004.

[82] Roli F., Giacinto G. (2002), Design Of Multiple Classifier Systems, S. 199-226 in: Bunke H., Kandel A. (Hrsgg.) (2002), Hybrid Methods in Pattern Recognition, World Scientific Publishing Co. Singapore 2002, http://citeseer.ist.psu.edu/552125.html : 26.9.2004, S. 223f. Das zitierte Kapitel gibt einen hervorragenden Überblick über das Design von Hybridklassifikatoren.

[83] Ein Klassifikator (*classifier*) ist eine Rechenvorschrift (Algorithmus) für die Zuordnung einer Klasse.

[84] Jain, A. K., Duin, R. P., Mao, J. (2000), Statistical pattern recognition: A review, S. 4-37 in: IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(1).

Effekt zu vermeiden, sollte jeder Klassifikator für sich arbeiten und eine Klasse vorschlagen. Erst danach wird aus den Ergebnissen z.B. durch Abstimmen ein gemeinsames Ergebnis ermittelt.[85]

5.3. Von Wörtern und *n*-Grammen

Ein Wort-*n*-Gramm ist eine Kette von *n* Wörtern, ein Buchstaben-*n*-Gramm die Aufeinanderfolge von *n* Buchstaben und *n* aufeinander folgende Zeichen bilden ein Zeichen-*n*-Gramm.[86] Ein Text wird also nicht mehr entlang der Wortgrenzen zerlegt, sondern entlang willkürlich nach *n* Zeichen gesetzter Grenzen. Je kleiner man *n* wählt, umso mehr (aber wenig abwechslungsreiche) Sequenzen entstehen. Die im nächsten Punkt behandelte Ähnlichkeit kann sich also nicht nur auf Wörter, sondern auch auf andere lexikalische Phänomene beziehen wie eben *n*-Gramme.[87]

Um die statistischen Schwierigkeiten mit Wörtern zu lösen, zieht man darum in der Linguistik und der Computerlinguistik **Zeichen-*n*-Gramme** heran. Man wendet also den Kunstgriff an, das Ganze in seine Elemente zu zerlegen wie eine Hängebrücke in einzelne Taufasern. So erhält man bei gleicher Textlänge größere Datenmengen und kommt eher zu statistisch aussagekräftigen Untersuchungsergebnissen.[88] Zusätzlich zerlegt man einen Text nicht nur einmal in Zeichenabschnitte wie beim Zerschneiden eines gedruckten Textes in einzelne Schnipsel, sondern erzeugt alle möglichen Zerlegungen des Textes. Jedes einzelne Zeichen taucht dann mehrfach in verschiedenen *n*-Grammen auf. Das Zeichen § wird bei einer Zerlegung des Textes in 3-Gramme[89] einmal am Anfang, einmal in der Mitte und einmal am Ende eines 3-Gramms vorkommen.[90]

Um das *sparse-data* Problem zu umgehen, werden daher die Vorhersagen der langen und validen *n*-Gramme mit denen der kurzen und statistisch interessanten *n*-Gramme zu einem einzigen neuen Algorithmus kombiniert. Das Verhältnis, mit dem jede *n*-Gramm-Kategorie in das Ergebnis eingeht, kann zum einen gleichbleibend festgelegt werden, z.B. von vornherein auf einen festen Wert wie 0,5 eingestellt werden oder durch einen Erwartungsmaximierungsalgorithmus (*expectation maximization*, EM) für alle Vorhersagen berechnet werden (lineare Interpolation). Andererseits kann man sog. *back-off* Modelle verwenden, bei denen eine Berechnung mit langen und validen *n*-Grammen versucht wird und wenn diese statistisch nicht zufriedenstellend ist, weil das betreffende *n*-Gramm seltener als ein bestimmter Schwellenwert auftaucht, dann wird auf kürzere *n*-Gramme zurückgegriffen.[91]

---

[85] Die unternommenen Vorversuche mit kombinierten Klassifikatoren lassen keine Systematik erkennen und werden hier nicht wiedergegeben, weil sie bei nur 140 Dokumenten statistisch nicht aussagekräftig sind.

[86] Die Buchstaben 3-Gramme von ‚Fachgebiet' sind Fac, ach, chg, hge, geb, ebi, bie, iet und et_.

[87] Zu *n*-Grammen siehe bei Fußn. 86. Damashek, M. (1995), Gauging similarity via n-grams: Language-independent sorting, categorization, and retrieval of text, S. 843-848 in: Science 1995: 267 vergleicht die beiden Verfahren und kommt zu dem Schluss, dass *n*-Gramme dem Wortansatz ohne linguistische Zusatzinformation meist überlegen sind. Cowie, J., Ludovik, E., Zacharski, R. (1998), An autonomous, web-based multilingual corpus collection tool, S. 142-148 in: Proceedings of the International Conference on Natural Language Processing and Industrial Applications (NLP+IA), University of Moncton Moncton 1998 sehen den Grund dafür darin, dass *n*-Gramme Wortstamm, -beugung und -zusammensetzung als ähnlich erkennen. Möglicherweise kann man sich mit *n*-Grammen auch an die optimale Menge vergleichbarer Textteilchen annähern. Es wäre interessant, den Zusammenhang zwischen der sprachlichen Dichte einer Schrift und dem besten *n* für *n*-Gramme zu untersuchen.

[88] Voraussetzung für die Übertragbarkeit der Ergebnisse ist allerdings die Validität der Indikatoren, also dass sich im Hinblick auf die zu untersuchende Eigenschaft einzelne Fasern wie die gesamte Hängebrücke verhalten oder *n*-Gramme wie Wörter.

[89] *N*-Gramme der Länge 2 werden Bigramme (*bigrams*) genannt, *n*-Gramme der Länge 3 Trigramme (*trigrams*).

[90] Werkzeuge zur Zerlegung von Texten in *n*-Gramme sind im Internet frei erhältlich, z.B. unter http://odur.let.rug.nl/~vannoord/TextCat/: 10.10.2003. Die meisten dieser Programme wurden für die Sprachenidentifizierung geschrieben, können aber auch für andere Aufgaben verwendet werden.

[91] Zur Kombination verschiedener *n*-Gramm Modelle ausführlicher Manning C.D., Schütze H. (1999), Foundations of Statistical Natural Language Processing, MIT Press Cambridge (USA) und London 1999, S. 217-225.

Schließlich gibt es noch *skipping n*-Gramme oder ***s*-Gramme**[92]. Bei der Erzeugung von *s*-Grammen werden nicht aufeinander folgende Zeichen verkettet, sondern ein oder mehrere Zeichen ausgelassen (*skipped*). Die Folge ist, dass bestimmte Buchstabenkombinationen, die in direkter Abfolge nicht möglich sind (in kaum einer Sprache ist die Aufeinanderfolge von vier Mal demselben Zeichen möglich), als *s*-Gramm auftaucht (das Wort „Beerengelee“ wird u.a. zu „ee_ _ _ _ _ _ee“). Daher können mehr *n*-Gramme der gleichen Länge erzeugt werden. Alternativ kann man *n* größer wählen, so dass von vornherein mehr *n*-Gramme erzeugt werden. Abgesehen vom statistischen Effekt durch die höhere Anzahl gibt es bisher noch keine gesicherten Erkenntnisse, welche Vorzüge oder Nachteile *s*-Gramme gegenüber *n*-Grammen haben.

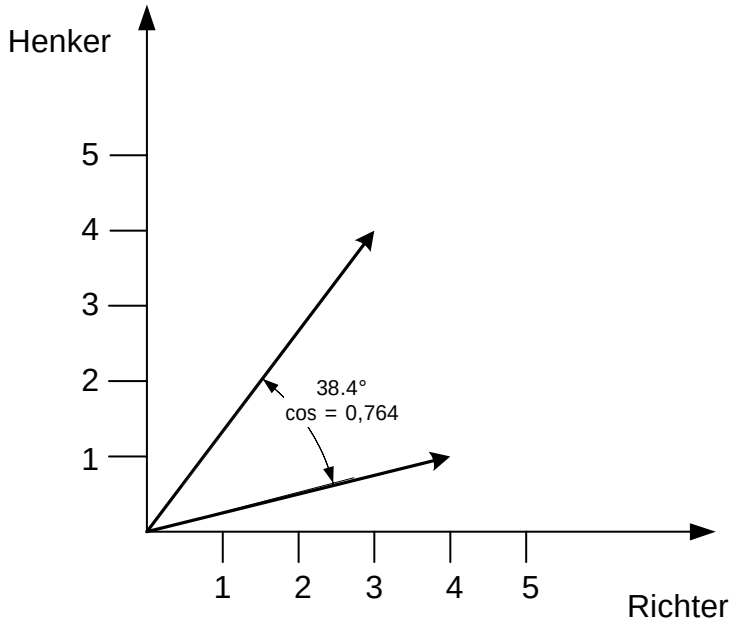
Eine weitere Aufbereitung der Frequenzdaten wird mit **Hapaxlegomena** gemacht. Hapaxlegomena sind Wörter, die in einer Textsammlung nur einmal vorkommen,[93] und die oft aus den Frequenztabellen getilgt werden, weil man sie für insignifikant hält.

### 5.4. Kosinusähnlichkeit der Merkmalsvektoren

Um die Ähnlichkeit zu berechnen, wird für jedes Objekt ein Merkmalsvektor erstellt. Die Merkmale eines Objekts sind die Frequenzen seiner Bestandteile, bei einem Text etwa die enthaltenen Wörter. Die vorkommenden Wörter und ihre Anzahl werden in eine Frequenztabelle eingetragen, die man auch als Vektor darstellen kann.

Eine Frequenztabelle mit 2 Wörtern wäre dann ein Vektor im zweidimensionalen Raum. Als Beispiel werden zwei Texte nur auf das Vorkommen der zwei Wörter ‚Richter’ und Henker’ untersucht. Die Frequenz von ‚Richter’ wird auf der x-Achse und die von ‚Henker’ auf der y-Achse aufgetragen. Für den ersten Text (Richter 3, Henker 4) ergibt sich ein steiler Vektor, für den zweiten (Richter 4, Henker 1) ein flacher (Grafik 1).

**Grafik 1: Textvergleich durch den Kosinus ihrer Vektoren**




[92] Der Ausdruck *s-gram* wurde geprägt von Pirkola, A., Keskustalo, H., Leppänen, E., Känsälä, H., Järvelin, K. (2002), Targeted s-gram matching: a novel n-gram matching technique for cross-and monolingual word form variants, in: Information Research 7 (2002): 2,
http://InformationR.net/ir/7-2/paper126.html : 31.3.2004.

[93] Unter http://www.nicolaas.net/hapax/index.php? : 2.4.2004 kann man sich aus einem eigenen Text alle Hapaxlegomena heraussuchen lassen.

Als Maß für die Ähnlichkeit der Vektoren könnte man den geometrischen Abstand der Vektorenendpunkte berechnen. Das ist im zweidimensionalen Raum noch relativ einfach, wird aber mit Zunahme der Dimensionen (jede gemessene Frequenz ergibt eine neue Dimension) zunehmend aufwändig.

Zur Vereinfachung der Berechnung kann der Kosinuswert des Winkels zwischen den beiden Vektoren genommen werden. Der Kosinus des Winkels ist 1 wenn der Winkel 0 Grad hat. Bei einem größeren Winkel nähert sich der Kosinus dem Wert 0, der minimalen Ähnlichkeit, wenn die Frequenztabellen keine gemeinsame Dimension haben bzw. die Texte kein Wort gemeinsam. Ein Argument für die Verwendung des Kosinus statt Winkelgraden ist, dass spitze bis rechte Winkel negative Werte annehmen können, der Kosinus aber immer größer oder gleich Null ist. Anders ausgedrückt: Der Kosinus nimmt nie einen negativen Wert an, und die Ähnlichkeit bewegt sich damit immer zwischen 0 (gar nicht ähnlich) und 1 (völlig übereinstimmend), wodurch man die Werte auch intuitiv gut vergleichen kann.

Wenn nun einen dritten Text konstruieren würde, der dreimal den Text 2 enthält, dann würde das Wort ‚Richter' zwölfmal und ‚Henker dreimal vorkommen. Der Vektor des dritten Textes würde genau auf dem Vektor des zweiten zu liegen kommen. Der Winkel zwischen ihnen wäre Null, der Kosinus eins und sie würden damit dieselbe Klasse zugewiesen bekommen. Nach geometrischem Abstand liegen die Endpunkte der Vektoren zu Text 1 und Text 2 aber viel näher als einer von ihnen zu dem langen Vektor zu Text 3. Das bedeutet, dass die Länge von Vektoren durch den Gebrauch des Winkels (Kosinus) nicht mehr ins Gewicht fällt. Mit anderen Worten kommt es nicht mehr auf die absolute, sondern auf die relative Frequenz an. Damit können Vektoren unterschiedlicher Länge, bzw. lange mit kurzen Texten vergleichen. Trotz der beiden erheblichen Vereinfachungen würde die Berechnung des Winkels zu jedem anderen Vektor mit steigender Anzahl klassifizierter Objekte bald zu lange dauern. Hier bietet das *Nearest Neighbour*-Verfahren einen Ausweg.

### 5.5. *Nearest Neighbour*-Verfahren

Die Klassifikation durch den ‚Nächsten Nachbarn' (NN)[94] beruht auf der Annahme, dass ähnliche Dokumente der gleichen Klasse angehören. Um ein Dokument zu klassifizieren, sucht der Klassifikator das ähnlichste Dokument und weist dessen Klasse dem zu klassifizierenden Dokument zu.

Identische Dokumente erzeugen identische Vektoren, die den Abstand Null haben, so dass jedes identische Dokument richtig klassifiziert wird. Wenn ein Dokument aber am Rand der Klasse liegt, kann der nächste Nachbar bereits eine andere Klasse haben. Um diesen Fehler zu vermeiden, muss man Information über die Lage der umliegenden Dokumente derselben Klasse in die Heuristik einbauen. Der *k*-NN-Klassifikator[95] bezieht daher *k* weitere Dokumente derselben Klasse in seine Ähnlichkeitsberechnung mit ein, die dann durch Abstimmen (*voting*) über die Klasse entscheiden. Je mehr nahe Nachbarn ihre Information in die Abstimmung einbringen können, umso besser wird das Ergebnis, aber je mehr ferne Nachbarn einbezogen werden, umso schlechter wird es. Meistens wird *k* daher zwischen 3 und 20 gewählt.

Das Verfahren bringt bereits bei einem klassifizierten Dokument ein Ergebnis. Es hat eine steile Lernkurve, verbessert sich also schnell mit zunehmenden Lerndaten. Das bedeutet für den Korpusneuaufbau, dass bereits ab dem zweiten Dokument der Übergang zur automatischen Erkennung beginnt,[96] weil bereits eine Klassifikation vorgeschlagen wird. Die intellektuelle Korrektur der Klasse

---

[94] Englisch: *Nearest Neighbour*, *Nearest Neighbor*, *Next Neighbor*, *Next Neighbour classifier*.

[95] Manning C. D., Schütze H. (1999), Foundations of Statistical Natural Language Processing, MIT Press, Cambridge, London, 1999, Kapitel 16.4, S. 604ff.

[96] Day, D., Aberdeen, J., Hirschman, L., Kozierok, R., Robinson, P., Vilain, M. (1997), Mixed-initiative development of language processing systems, S.348-355 in: 5th Conference on Applied Natural Language Processing (ANLP 97), Association for Computational Linguistics Washington D.C. 1997, http://acl.ldc.upenn.edu/A/A97/A97-1051.pdf : 21.9.2004.

(semi-automatische Klassifizierung) wird immer seltener nötig und schließlich sollte die Fehlerschwelle so niedrig werden, dass die Klassifizierung unbeaufsichtigt vonstatten gehen kann. Während statistische Verfahren mit steigenden Lerndaten immer schlechter korrigiert werden können[97] und im Extremfall gar nicht mehr lernen,[98] ist das bei NN-Verfahren nicht der Fall, weil jedes Dokument einen eigenen Ort hat[99] und als eines von wenigen Dokumenten (nicht als eine statistische Nichtigkeit unter vielen) entscheidend sein kann.

Von großem Einfluss ist auch das Ähnlichkeitsmaß, das der Berechnung des ‚nächsten' Nachbardokuments zugrunde gelegt wird.[100]

5.6. Ähnlichkeitsmaß und *k*-NN-Algorithmus

Kaum ein Klassifizierungsverfahren wendet das NN-Verfahren in Reinform an, weil man dazu ja den Abstand eines Merkmalsvektors zu allen Tausend oder Millionen anderen berechnen müsste. Daher werden die zu vergleichenden Dokumente oft vorher schon reduziert, möglichst ohne die nächsten Nachbarn dabei zu verlieren, und dann wird das NN-Verfahren angewandt.[101] Wenn die Reduktion allerdings zu stark ist, dann wird das ganze NN-Verfahren wertlos.

Wegen der z.T. sehr kurzen und nicht als Einzelwort vorliegenden Informationen (Verweisadresse) wird das Ähnlichkeitsmaß durch 3- und 4-Gramm-Ähnlichkeit berechnet. Aus Termfrequenz[102] (TF) und Dokumentfrequenz[103] (DF) ergibt sich nach der anerkannten Formel TF.IDF=TF/DF die Relevanz eines *n*-Gramms für die Klassifizierungsaufgabe. Das bedeutet, dass ein *n*-Gramm umso mehr über die Ähnlichkeit zweier Dokumente aussagt, je öfter es in einem Dokument und je seltener es in allen Dokumenten vorkommt.

Nun wird der Einfluss unterschiedlicher Textmengen auf die Frequenzen ausgeglichen, indem die TF eines *n*-Gramms mit der TF des häufigsten *n*-Gramms in Beziehung gesetzt wird:

$$TF_{neu} = 0{,}5 + 0{,}5 \bullet \frac{TF}{TF_{\max}}$$

Streiter O. (2001), Corpus-based parsing and treebank development, S. 115-120 in: ICCPOL 2001, 19th International Conference on Computer Processing of Oriental Languages, Seoul (Korea) 2001.

[97] Day et al. (1997) a.a.O.

[98] Kurohashi, S., Nagao, M. (1998), Building a Japanese parsed corpus while improving the parsing system, S. 719-724 in: Rubio A., Gallardo N., Castro R., Tejada A. (Hrsgg.) (1998), First International Conference on Language Resources & Evaluation, Granada (Spanien) 1998.

[99] Aha, D. W. (1997), Editorial-lazy learning, S. 1-3 in: Artificial Intelligence Review 1997: 11.

[100] Eine nützliche Einführung in die verschiedenen Methoden der Dokumentklassifizierung bieten Han, E.-H. S., Karypis, G. (2000), Centroid-based document classification: Analysis & experimental results, S. 424-431 in: Zighed, D. A., Komorowski, J., Zytkow, J. (Hrsgg.) (2000), Principles of Data Mining and Knowledge Discovery, Proceedings of the 4th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD) 2000, Reihe: Lecture Notes in Computer Science; Lecture Notes in Artificial Intelligence Vol. 1910, Lyon (Frankreich) 2000, http://www.cs.umn.edu/karypis: 18.6.2004.

[101] Mit dem Laufzeitproblem bei der exakten *k*-Nächste-Nachbarn-Suche beschäftigt sich die digitale Dissertation (FU Berlin) von Heinrich-Litan L. N. (2002), Exakte L∞-Nächster-Nachbar-Suche in hohen Dimensionen - Exact L∞-Nearest-Neighbor Search in High Dimensions, http://www.diss.fu-berlin.de/2003/80/index.html : 25.9.2004, in der nicht nur die Formel zur Laufzeitberechnung in Abhängigkeit von der Dokumentanzahl und den Dimensionen angegeben wird, sondern auch eine Methode für einen optimalen Ausgleich zwischen Laufzeit und Speicherbedarf.

[102] Vorkommen des *n*-Gramms in einem Dokument.

[103] Vorkommen des *n*-Gramms in allen Dokumenten.

Der neue TF.IDF berechnet sich dann: $\frac{TF}{IDF_{neu}} = \frac{TF_{neu}}{DF}$

Durch eine Kosinusnormalisierung wird erreicht, dass die unterschiedliche Textmenge der Dokumente nicht ins Gewicht fällt:

$$\frac{TF}{IDF_{normal}} = \frac{\frac{TF}{IDF_{neu}}}{\sqrt{\sum \frac{TF}{IDF_{neu\,2}}}} = atc - weight$$

Das so berechnete Ähnlichkeitsmaß heißt *atc-weight* oder ATC-Gewicht.[104]

Der NN-Algorithmus holt sich in einem ersten Schritt alle Dokumente, die die relevantesten *n*-Gramme des zu klassifizierenden Dokuments enthalten. Nur diese Auswahl wird dann auf die Kosinusähnlichkeit (ATC-Gewicht) aller *n*-Gramme untersucht.[105] Nur die besten NN dürfen dann mit ihrer Klasse über die Klasse des zu klassifizierenden Dokuments abstimmen.

## 6. Experimente

### 6.1. Versuchsbeschreibung

Zum Testen wurden 140 Dokumente intellektuell klassifiziert und gelten als ideale Klassifizierung. Dann versucht jeder der neun Klassifikatoren (siehe Tabelle 4), dieses Ergebnis zu reproduzieren. Das erste Dokument kann mangels klassifizierter Nachbarn nicht klassifiziert werden. Sodann kommt dieses erste Dokument mit der Information über seine Klassen zu den Lerndaten, aus denen der Nachbar gewählt wird. Der Klassifikator rät für das zweite Dokument zwangsläufig die Klassen des bisher einzig klassifizierten Dokuments. Ab dem dritten Dokument sollte der Lerneffekt einsetzen. Um die Größe des Lerneffekts abschätzen zu können, wird ein zehnter Klassifikator eingesetzt, der einen zufälligen Nachbarn auswählt (*random NN*).

Die X-Achse der Grafiken 2 bis 6 zeigt die Anzahl der klassifizierten Dokumente, also der möglichen Nachbarn. Das Dokument 101 wird aufgrund der Informationen der 100 bis dahin klassifizierten Dokumente klassifiziert. Auf der Y-Achse ist jeweils die Übereinstimmung des Klassifikators mit der intellektuellen Klassifikation aufgetragen. Ein Wert von 0.5 bedeutet, dass jedes zweite Dokument richtig klassifiziert werden konnte. Je nach Aufgabe und Anzahl der Klassen kann das ein gutes oder schlechtes Ergebnis sein. Nicht korrekte Klassifizierungen können an der Wahl eines falschen NN oder am Fehlen eines NN der richtigen Klasse liegen. Erst wenn jede Klasse zumindest einen NN aufweist, kann ein Klassifikator wenigstens theoretisch stets die richtige Klasse vorschlagen; vorher ‚kennt' er die Klasse überhaupt nicht.[106] Dieser Effekt, der je nach Anzahl der Klassen länger oder kürzer anhält, schlägt sich aber auch auf die Ergebnisse des Zufallsklassifikators nieder, so dass dieser als Vergleichsmaßstab herangezogen werden kann. Von Klassifikatoren, die sich nach 140 Trainingsdokumenten noch nicht vom Zufallsklassifikator absetzen konnten, kann man annehmen, dass sie für die Klassifizierung dieser Klasse ungeeignet sind.

Für jedes richtige Ergebnis wird der Wert 1 und für jedes falsche der Wert 0 vergeben. Durch Extrapolierung erhält man eine Kurve (Siehe Grafiken 2-6). Hier werden die 20 letzten Ergebnisse

---

[104] Weitere Ähnlichkeitsmaße werden im Kapitel 4 über die Termextraktion bei den Fußn. 204 ff. besprochen.

[105] Manning, C. D., Schütze, H. (1999), Foundations of Statistical Natural Language Processing, MIT Press, Cambridge, London, 1999, Kapitel 8.5, S. 294ff, http://www-nlp.stanford.edu/fsnlp/ : 18.6.2004.

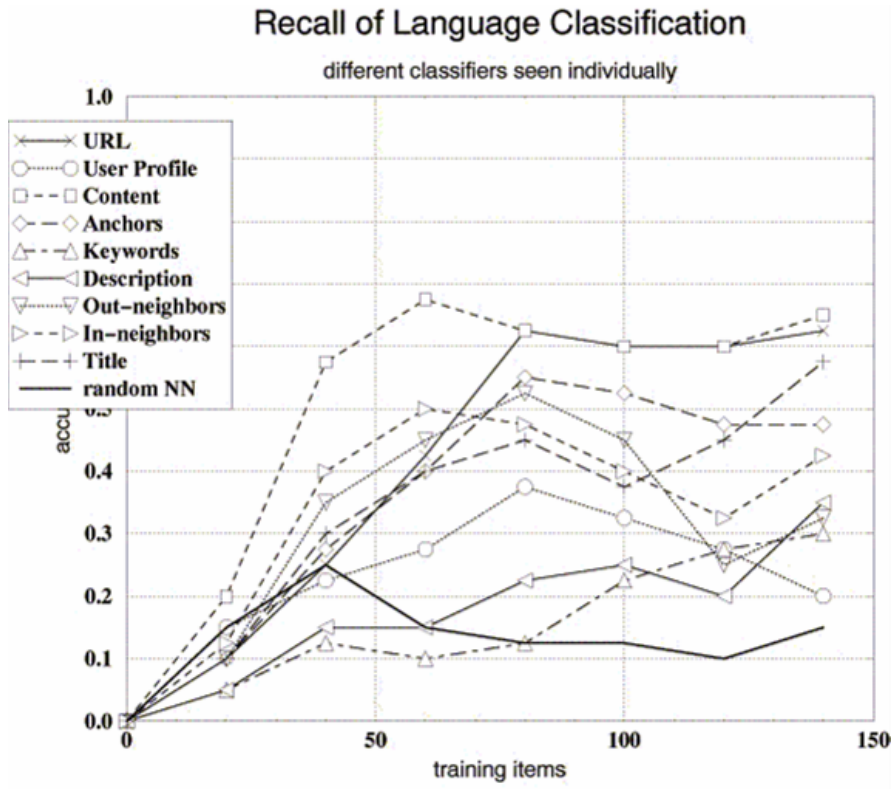
[106] Für den Klassifikator kommt mit jedem Beispieldokument, das eine bisher unbekannte Klasse trägt, eine neue Klassifizierungsmöglichkeit hinzu. Das Hinzufügen einer konzeptuell neuen Klasse zu einem späteren Zeitpunkt ist für diesen Klassifikator also nichts Außergewöhnliches und seine Ergebnisse werden deshalb nur vorübergehend etwas schlechter.

aufaddiert, die Graphen zeigen aber noch deutlich die Schwankungen der binären Ereignisse. Der Endpunkt der Graphen nach 140 Dokumenten bezeichnet also nicht den durchschnittlichen Erfolg bei allen Versuchen, sondern in etwa die Erfolgswahrscheinlichkeit für den unmittelbar folgenden Versuch. Aus dem Schwanken der Graphen kann man ablesen, dass der Endpunkt auch nicht die Güte der einzelnen Klassifikatoren endgültig klärt, weil sich bei Fortführung des Experiments über weitere 100 oder 10.000 Dokumente ein anderer Klassifikator durchsetzen kann.

6.2. Versuchsergebnisse

Die **Sprachenidentifizierung** ist für einsprachige Texte bei hinreichend viel Untersuchungsmaterial und hinreichender Unterschiedlichkeit der zu erkennenden Sprachen ein gelöstes Problem.[107] Hier waren allerdings auch mehrsprachige Dokumente zu erkennen.

**Grafik 2: Sprachklassifikation**

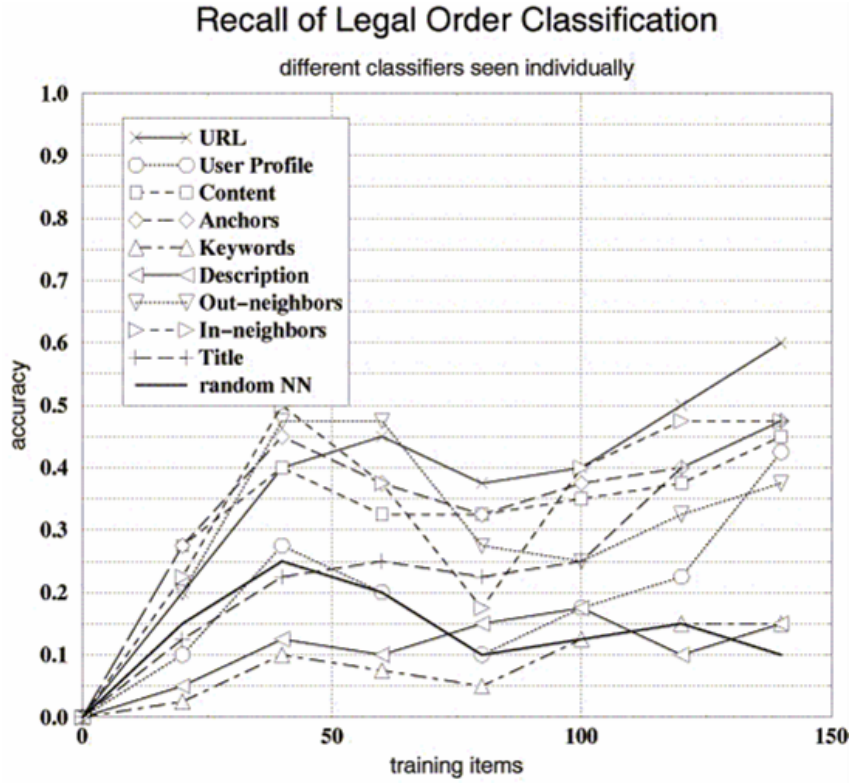



Der Inhalt zeigt sich von Anfang an als bester Klassifikator (als ‚content' in Grafik 2 gekennzeichnet) und erreicht eine 60 bis 70-prozentige Wahrscheinlichkeit, die richtige Sprache oder Sprachkombination zu erreichen. Auch die URL erzielt mit dem Textstück „.it/" oder „.de/" hervorragende Ergebnisse. Die Vorhersagen der Sprache mit dem Mitarbeiterprofil waren sehr schlecht, was wohl darauf zurückführen ist, dass alle Mitarbeiter mehrsprachig sind und für alle Sprachen gleich viel Material zur Arbeit benötigen. Die doch enttäuschenden Ergebnisse in dieser Sparte ergeben sich durch die englischsprachigen Einschübe vor allem im Wirtschafts- und Bankenrecht. Auch wenn ein Dokument nicht einmal ganze Sätze, sondern nur einige Fachausdrücke in englischer Übersetzung enthält, könnte man dieses Dokument zur Suche eines englischsprachigen Beg-

[107] „Die automatische Sprachenidentifizierung für elektronische Dokumente, deren Mindestlänge eine bestimmte Wortzahl überschreitet und die regulären Text enthalten, kann als weitgehend gelöstes Problem betrachtet werden." Langer, S. (2002): Grenzen der Sprachenidentifizierung, S. 99-106 in: Tagungsband KONVENS 2002, DFKI GmbH Saarbrücken 2002, S. 99, http://konvens2002.dfki.de/cd/pdf/19V-langer.pdf : 27.9.2004. Nicht immer bekommt man regulären Text im Internet, weil Inhalte oft in Fenster mit anderen Inhalten (Werbung, Bilder, Navigationshilfen, Menüleisten) eingebunden werden und weil der übersandte Quelltext (anders als die Anzeige in einem Browser) oft mit Formatierungshinweisen (statt Inhaltshinweisen, die das Formatieren dem Browser überließen) übersäht ist.

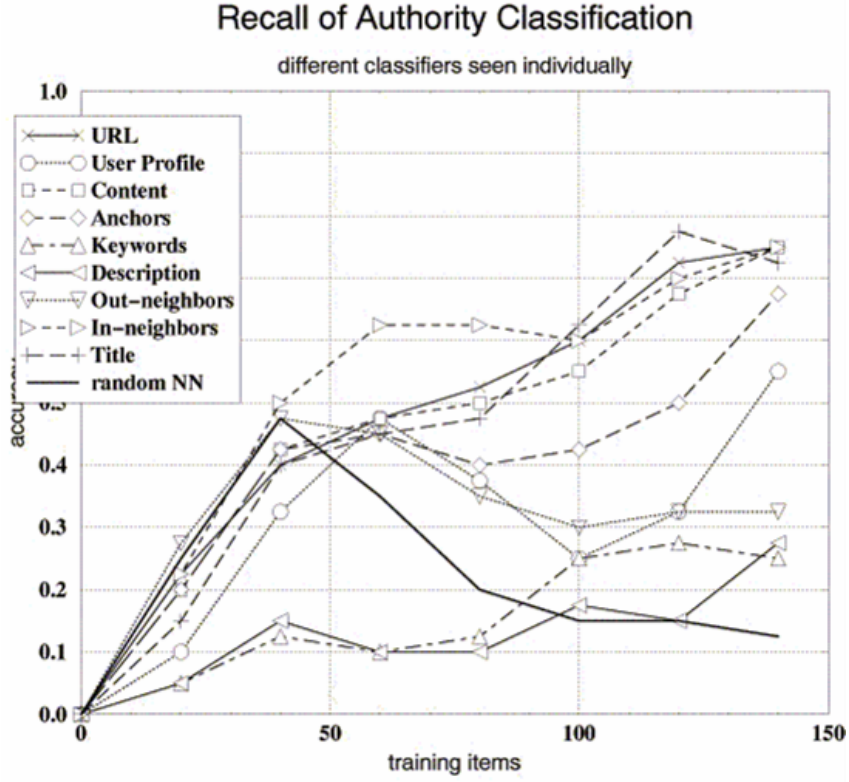
riffs einsetzen. Es soll daher nicht nur als einsprachig, sondern auch als Englisch klassifiziert werden, was für den Klassifikator eine eigene, neue Klasse ist. Insofern sind die Ergebnisse noch im Rahmen, wenn auch schlechter als bei einer originären Sprachenidentifizierung eines Textes. Die Methode eignet sich also nur in den seltenen Fällen, in denen Text nicht als solcher erkannt werden kann, weil er als Bild eingefügt wurde.

**Grafik 3: Klassifizierung der Rechtsordnung**



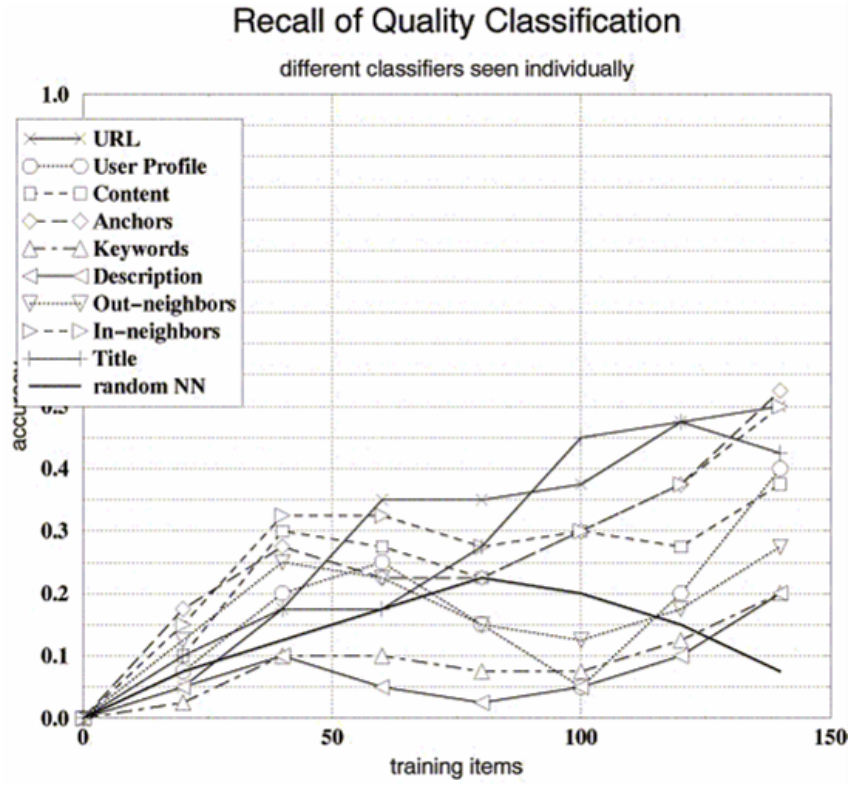

Bei der Dokumentklassifikation nach der **Rechtsordnung** erlaubt die URL mit den Anhängseln „.ch“, „.eu“ oder „.at“ eine gute Orientierung. Der URL- Klassifikator ist daher mit 60 % der beste zur Vorhersage der Rechtsordnung (Grafik 3). Die Metadaten sind zu selten, um konkurrenzfähig zu sein, was umso interessanter ist, weil sie gerade für solche schwer aus dem Inhalt erkennbaren Informationen (deutsches oder österreichisches Einkommensteuergesetz?) ersonnen wurden. Selbst wenn die Identifizierung der Klassen fehlerfrei wäre, müsste das Gros der Dokumente stets durch einen anderen Klassifikator erkannt werden. Dies gilt für alle zu erkennenden Klassen. Allerdings könnten Metadaten in Kombination mit einem anderen Klassifikator die Ergebnisse verbessern.

**Grafik 4: Klassifizierung der Norm- und Staatsorganisationshierarchie**




Bei der Klassifizierung der **Hierarchieebene** funktionieren Titel, URL, Inhalt und URL-Verweis (hinverweisende Links) etwa gleich gut; knapp dahinter liegen Verweise und Mitarbeiterprofil (siehe Grafik 4). Unerwartet ist, dass die hinverweisenden Links so viel besser abschneiden als die wegverweisenden. Eine Erklärung dafür ist, dass die Dokumente, die mit einer besonderen Klasse versehen werden müssen, stets von offiziellen Servern geladen werden. Diese seriösen Server bieten meist Text und einen Verweis auf die Wurzel oder die zentrale Seite. Andere Verweise werden hingegen nicht auf den Dokumentseiten, sondern in einer separaten Linkseite angeboten. Wegverweisende Links sind daher sehr selten und bieten zu wenig Information. Der Servername und evtl. der Ordner ist hingegen oft offiziell und enthält charakteristische bedeutungstragende Bruchstücke wie „Bundes“ oder „gesetz“ und erlaubt eine gute Klassifizierung.

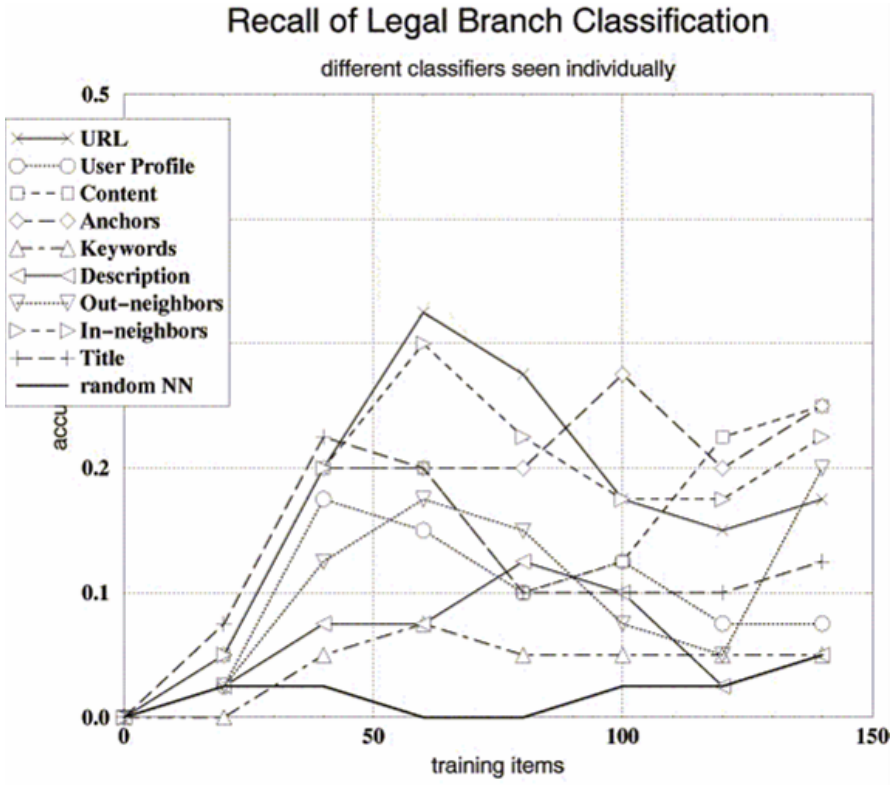
**Grafik 5: Klassifizierung der Rechtsqualität**




Metadaten und der Unterschied zwischen hin- und wegverweisenden Links wurde bereits besprochen. Insgesamt sind die Lernkurven bei der Klassifizierung der **Rechtsqualität** in Grafik 5 flach; nach 70 Dokumenten ist praktisch nur die URL deutlich besser als die Zufallsklassifikation. Bei Beendigung der Versuche stiegen acht von neun Graphen noch weiter an. Daraus kann man vielleicht generell den Schluss ziehen, dass unabhängig von der Anzahl der Klassen Versuche mit sehr viel mehr Dokumenten nötig sind, um zu einer verallgemeinerbaren Aussage gelangen zu können. Die hohe Güte heutiger Sprachenidentifizierer macht Hoffnung für die Erkennung von anderen Klassen, sofern die hierfür nötige Menge Lernmaterial oder hinreichend große Textmengen je Dokument erreicht werden.

Die Hypothese, dass das Mitarbeiterprofil bei der **Fachgebietsklassifikation** am besten abschneiden würde, bewahrheitete sich wohl deshalb nicht (siehe Grafik 6) weil die Fachgebiete kleiner geschnitten sind als das Arbeitsgebiet eines Mitarbeiters. Ins Arbeitsgebiet Verwaltungsrecht fällt etwa das besondere Verwaltungsrecht, das Baurecht, das Polizeirecht und das Verwaltungsprozessrecht. Außerdem müssen Mitarbeiter zur Bearbeitung eines Gebiets auch die Bedeutung eines Begriffs in anderen Gebieten zumindest eruieren.

**Grafik 6: Fachgebietsklassifikation**




Man beachte, dass die Dimensionen in Grafik 6 ganz andere sind als in den Grafiken 2 bis 5, bei denen die besten Klassifizierer stets über 0,5 lagen. Bei der Fachgebietsklassifikation liegen die besten Ergebnisse um die 20 %. Der Verlauf der Kurven in Grafik 6 wirkt deshalb besonders unstet. Die Lernkurven haben bei Abbruch des Experiments nach 140 Dokumenten aber keine eindeutig steigende Tendenz, wenn man die Endpunkte etwa mit den Datenpunkten nach 40 Dokumenten vergleicht. Wenn man als Kriterium für die praktische Brauchbarkeit postuliert, dass ein Klassifikationsvorschlag wenigstens öfter richtig sein sollte als falsch, dann deuten die Daten darauf hin, dass mit der hier vorgeschlagenen Methode auch mit wesentlich mehr Dokumenten keine brauchbare Fachgebietsklassifikation zu erreichen sein wird. Daher werden im Kapitel 2 zwei weitere Methoden zur Lösung dieses Problems untersucht.

6.3. Bewertung der Ergebnisse von NN-Klassifikatoren

Es wurde erwartet, dass die Präzision der Klassifikatoren von der Anzahl der Klassen abhängt, dass also die Norm- und Staatsorganisationshierarchie mit nur fünf Klassen acht mal leichter zu klassifizieren sein würde als das Fachgebiet mit 41 Klassen. Der Unterschied hat aber nur den Faktor drei. Außerdem wurde erwartet, dass 140 Dokumente leicht ausreichen würden, um fünf Klassen mit guten NN zu belegen, während die Lernkurve bei vierzig Klassen noch weiter nach oben zeigen würde. Die Lernkurven zeigen aber, dass Klassifikatoren selbst mit einer geringen Anzahl Lerndokumenten schon gute Arbeit leisten.

Ein Grund kann sein, dass die Klassifikatoren nicht aus dem gesamten Klassifikationsraum auswählen, weil sie ihn noch gar nicht ganz kennen. Das entspricht der Information, dass häufigere Klassen früher im Klassifikationsraum auftauchen und dass die Dokumente ungleichmäßig auf diesen Raum verteilt sind. Tatsächlich waren die Klassen sehr ungleich verteilt, z.B. waren die Sprachen Deutsch und Italienisch stark überrepräsentiert. Um diesen Vorteil auszugleichen, müsste ein Korpus zur Verfügung stehen, in dem alle Klassen in etwa gleich belegt sind.

## 7. Mögliche Verbesserungen

Die in Punkt 5.2. angesprochenen hybriden Klassifikatorsysteme könnten die Ergebnisse sicher verbessern. Bereits angesprochen wurde auch, den Linktext als Klassifikator zu verwenden.

Eine Möglichkeit zur Verbesserung könnte die Anwendung der Mengenlehre zur Reduzierung der Klassen sein. Ein zweisprachiges Dokument könnte als Dokument aufgefasst werden, das Eigenschaften zweier Mengen in sich vereint. Wenn man eine Sprache gut erkennt, dann steigen die Chancen, auch die zweite richtig zu erkennen – sowohl für den Algorithmus wie für den menschlichen Klassifizierer.

Weitere Untersuchungen des Einflusses der Textmenge je Dokument und der Trainingsmenge wären hilfreich. Wenn die relativ kleine Textmenge von Internetseiten die Qualität vermindert, dann wären die größeren Dokumente von Verlagen umso interessanter. Außerdem könnte man mehrere Dokumente, etwa vom gleichen Server, für die Klassifizierung zu einem gemeinsamen Dokument zusammenfassen.

Auch das Ausnutzen der bekannten Hierarchie- oder Nähebeziehung von einigen Fachgebieten könnte die Fachgebietsklassifizierung verbessern. Man könnte nur einfach zu erkennende Dokumente automatisch klassifizieren lassen und die schwierigen semiautomatisch oder ganz intellektuell klassifizieren. Das würde sich insbesondere bei der Sprachenidentifizierung anbieten. Dafür müsste aber ein Kriterium für einfach zu klassifizierende Dokumente gefunden werden, also für Dokumente, bei denen die Wahrscheinlichkeit eines Klassifizierungsfehlers gegen Null tendiert. Würde dies gelingen, dann könnte man das Kriterium in den Web-Robot integrieren und den Robot nur noch einfache Dokumente aus dem Web sammeln lassen. Möglicherweise müsste dann der Korpusaufbau semiautomatisch verlaufen: Ein Mitarbeiter würde die ersten fünf Dokumente intellektuell klassifizieren und der Web-Robot hätte dann Trainingsmaterial, anhand dessen er entscheiden könnte, welche weiteren Dokumente leicht zu erkennen und daher zu speichern sind. Wenn der Web-Robot dazu fähig ist, dann kann er das nächste Mal die gleiche Domain selbständig besuchen und nach geeignetem Material suchen. Damit würde sich das Korpus selbst mit klassifiziertem Material erneuern.

## 8. Zusammenfassung

In diesem Kapitel wurden einige Voraussetzungen für die erfolgreiche Rechtsterminologie erörtert und die Bedeutung von Korpora hervorgehoben. Ein Korpus ist über das kurzfristige Ziel und das einzelne Projekt hinaus verwendbar und erhaltenswert, stellt es doch Denken und Formulieren einer bestimmten Zeit und eines bestimmten Bereichs dar. Ein Korpus sollte daher auch öffentlich zugänglich sein.[108] Traditionelle Terminologie ist im Grunde genommen erst die Kondensierung und Veredelung dieses Wissens. Mit neuen Technologien kann ein Teil der Kondensierung maschinell ablaufen, beispielsweise die Generierung und Sortierung von Kontextstellen. Damit kann die intellektuelle Arbeit auf die wirklichen Probleme in der Rechtsvergleichung, Termdarstellung oder Nutzeranpassung konzentriert werden.

Es wurden die Quellen für ein Korpus diskutiert und die Bedeutung des Internets als Quelle hervorgehoben. Da der Nutzen eines Korpus von seiner sinnvollen Zusammenstellung aus gewichteten Klassen abhängt, wurden einige Klassen für mehrsprachige Rechtsdokumente vorgestellt. Dabei ist das in einem Dokument enthaltene rechtliche Wissen nicht das alleinige und vielleicht nicht einmal das hauptsächliche Kriterium, da eine ganze Reihe anderer Informationen als Klassifikatoren für eine semiautomatische oder später auch automatische Klassifikation in Frage kommen. Es empfehlen sich insbesondere der Dokumentinhalt, die Metadaten, (trotz schlechter Ergebnisse evtl. auch) die Mitarbeiterprofile und formale Kriterien als Klassifikatoren. Alle diese Informationen sind wichtig,

[108] Erst Recht dann, wenn es Minderheitensprachen wie hier Ladinisch enthält.

auch wenn nicht alle Dokumente diese Informationen tragen oder die vorhandenen nicht für eine Klassifizierung hinreichen. Für jedes Dokument sollten also alle Informationen gespeichert werden, weil beim raschen Fortschreiten der Methodenentwicklung früher oder später alle Daten eine sinnvolle Verwendung finden könnten.

Die übergreifende Erkenntnis hier ist, dass der Aufbau eines Korpus kein unüberwindliches Hindernis ist, wenn *Open Source*-Programme verwendet werden, auf Standards geachtet wird, allen Informationen Beachtung geschenkt wird und von einer semiautomatischen Klassifikation ausgehend der Automatisierungsgrad zunehmend erhöht wird.

Allerdings blieben die Ergebnisse der Fachgebietserkennung durchweg so schlecht, dass sie praktisch nicht zu gebrauchen ist. Daher soll im nächsten Kapitel das Problem der Fachgebietserkennung noch einmal speziell untersucht und mit zwei anderen Verfahren einer zufrieden stellenden Lösung zugeführt werden. Die hierzu notwendige zusätzliche Fachgebietsinformation über Texte werden die nach Fachgebieten eingeteilten Terme liefern.

# Literaturangaben zu Kapitel 1

Aha D.W. (1997), Editorial- lazy learning, S. 1-3 in: Artificial Intelligence Review 1997/11.

Banko M., Brill E. (2001), Scaling to Very Very Large Corpora for Naturale Language Disambiguation, S. 26-33 in: Meeting of the Association for Computational Linguistics 2001, http://citeseer.nj.nec.com/banko01scaling.html : 18.6.2004.

Brill E. (1995), Transformation-based error-driven learning and natural language processing: A case study in part of speech tagging, Computational Linguistics, 1995.

Brin, S., Page, L. (1998), The anatomy of a large-scale hypertextual Web search engine, S. 107-117 in: Computer Networks and ISDN Systems Vol. 30, Nr.1-7 1998, http://citeseer.nj.nec.com/brin98anatomy.html: 10.10.2003.

Bonnet E., Gaussier E., Langé J.-M. (1994), A method for automatic extraction of terms from bilingual corpora, in: Proceedings of the 14th International Conference on Artificial Intelligence KBS, Expert Systems and Natural Language (AVIGNON-94), EC2 Nanterre 1994.

Carl M., Schaible J., Pease C. (1998), Enhancing translation memory (TM) technologies with linguistic intelligence, MULTI-DOC Deliverable D 4.1 WP 6, Kommission der Europäischen Gemeinschaften Luxemburg 1998.

Chakrabarti S., Dom B. und Indyk P. (1998), Enhanced hypertext categorization using hyperlinks, S. 307-318 in: Haas L. M., Tiwary A. (Hrsgg.) (1998), Proceedings ACM SIGMOD International Conference on Management of Data, ACM Press Seattle 1998.

Charniak, E. (1996), Tree-bank grammars, S. 1031-1036 in: Proceedings of the 13th National Conference on Artificial Intelligence (AAAI-96), American Association for Artificial Intelligence (AAAI) Portland 1997.

Chien L.-F., Chen C.-L. (2001), Incremental extraction of domain-specific terms from online text resources, S. 89-109 in: Bourigault D., Jacquemin C. und L'Homme M.-C., (Hrsgg.) (2001), Recent Advances in Computational Terminology (Natural Language Processing), John Benjamins Amsterdam 2001.

Cowie, J., Ludovik, E., Zacharski, R. (1998), An autonomous, web-based multilingual corpus collection tool, S. 142-148 in: Proceedings of the International Conference on Natural Language Processing and Industrial Applications (NLP+IA), University of Moncton Moncton 1998.

Damashek M. (1995), Gaugin similarity via n-grams: Language-independent sorting, categorization, and retrieval of text, S. 843-848 in: Science, 1995: 267.

Day D., Aberdeen J., Hirschman L., Kozierok R., Robinson P., Vilain M. (1997), Mixed-initiative development of language processing systems, s. 348-355 in: 5th Conference on Applied Natural Language Processing (ANLP-97), Association for Computational Linguistics, Washington D.C. 1997.

De Carolis B., De Rosis F. und Pizzutilo S. (1997), Generating user-adapted hypermedia from discourse plans, in: Lenzerini M., (Hrsg.) (1997), Proceedings of the 5th Congress of the Italian Association for Artificial Intelligence (AI*IA97), Lecture Notes in Artificial Intelligence (LNAI) Teilreihe von Lecture Notes in Computer Science (LNCS), 1321, Springer Heidelberg 1997.

Engberg, J. (1993), Prinzipien einer Typologisierung juristischer Texte, S. 31-38 in: Fachsprache 15 1/2, Wien 1993.

Faber P., Lopés Rodriguez C. I. und Tercedor Sánchez M. I. (2002), Utilización de técnicas de corpus en la representatición del conocimiento médico, S. 167-197 in: Terminology, 2002, 7(2).

Find J., Kobsa A. und Nill A., User-oriented adaptivity and adaptability in the AVANTI project, in: Conference "Design for the Web: Empirical Studies, Microsoft, Redmond, WA, 1996.

Frilling S. (1994), Textsorten in juristischen Fachzeitschriften, Internationale Hochschulschriften 138, Waxmannverlag Münster/New York 1995, zugl. Diss. Univ. Münster 1994.

Furuse, O. and Iida, H. (1992), An example-based method for transfer-driven Machine Translation, in: The Third International Conference on Theoretical and Methodological Issues (TMI), Empiristic vs. Rationalist Methods in MT, Canadian Workplace Automation Research Center Montréal 1992.

Han E.-H. S. und Karypis G. (2000), Centroid-based document classification: Analysis & experimental results, S. 424-431 in: 2000, URL, http://www.cs.umn.edu/karypis: 10.10.2003.

Heinrich-Litan L. N. (2003), Exakte L∞-Nächster-Nachbar-Suche in hohen Dimensionen - Exact L∞-Nearest-Neighbor Search in High Dimensions, digitale Dissertation an der FU Berlin, http://www.diss.fu-berlin.de/2003/80/index.html : 25.9.2004.

Herberger M., Systematik des Bundesrechts, Projektbericht, Juristisches Internetprojekt Saarbrücken, 1997, http://www.jura.uni-sb.de/BGBI/BGBLSYST.HTML: 10.10.2003.

Jain A. K., Duin R. P. und Mao J. (2000), Statistical pattern recognition: A review, S. 4-37 in: IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(1).

Kang H.J., Doermann D. (2003), Evaluation of the Information Theoretic Construction of Multiple Classifier Systems, Proceedings of the 7th International Conference on Document Analysis and

Recognition (ICDAR 2003), IEEE Edinburgh 2003, http://citeseer.ist.psu.edu/ps/698712 : 26.9.2004.

Kurohashi, S., Nagao, M. (1998), Building a Japanese parsed corpus while improving the parsing system, S. 719-724 in: Rubio A., Gallardo N., Castro R., Tejada A. (Hrsgg.) (1998), First International Conference on Language Resources & Evaluation, Granada (Spanien) 1998.

Langer S. (2002), Grenzen der Sprachenidentifizierung, S. 99-106 in: Tagungsband KONVENS 2002, DFKI Saarbrücken 2002, http://konvens2002.dfki.de/cd/pdf/19V-langer.pdf: 27.9.2004.

Larkey L. S. und Croft W. B. (1996), Combining classifiers in text categorization, S. 289-297 in: Proceedings of SIGIR-96, 19. International Conference on Research and Development in Information Retrieval, ACM Press, New York, US, Zürich, CH, 1996.

Luckhardt H.-D., Harms I., Virtuelles Handbuch Informationswissenschaft, Universität des Saarlandes, http://www.is.uni-sb.de/studium/handbuch//exkurs.ind.php#intellind : 19.9.2004.

Lundquist, L. (1979), Teksttypbestemmelse af en lovtekst via en semantisk dybdestruktur, in: Linnarud, M., Svartvik, J. (Hrsgg.) (1979), Kommunikativ Kompetens och fackspråk, SyMPOSIUM Södertälje 1978, ASLA Uppsala 1979.

Manning C. D. und Schütze H. (1999), Foundations of Statistical Natural Language Processing, MIT Press Cambridge, London, 1999.

Mayer F. (2000), Terminographie im Recht: Probleme und Grenzen der Bozner Methode, S. 295-306 in: Veronesi, D., Rechtslinguistik des Deutschen und Italienischen, Unipress Padova 2000.

Nohr H. (2001), Automatische Indexierung, Einführung in betriebliche Verfahren, System und Anwendungen, Materialien zur Information und Dokumentation, Verlag für Berlin-Brandenburg, Potsdam, 2001.

Ott S. (2003), Linking und Framing: Ein Überblick über die Entwicklung im Jahre 2002, JurPC Web-Dok. 14/2003, http://www.jurpc.de/aufsatz/20030014.htm: 23.9.2003.

Rasmussen, K. W; Engberg, J. (1999), Genre Analysis of Legal Discourse, S. 113-132 in: Hermes - Journal of Linguistics 1999, 22.

Roli F., Giacinto G. (2002), Design Of Multiple Classifier Systems, S. 199-226 in: Bunke H., Kandel A. (Hrsgg.) (2002), Hybrid Methods in Pattern Recognition, World Scientific Publishing Co. Singapore 2002, http://citeseer.ist.psu.edu/552125.html : 26.9.2004.

Schmidt-Wigger A. (1998), Building consistent terminologies, Poster in: Proceedings of the First Workshop on Computational Terminology (COMPUTERM'98) at the 17th International

Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics (COLING-ACL'98), ACL University of Montreal (Hrsgg.) Montreal 1998,
http://www.iai.uni-sb.de/docs/term2.pdf : 1.10.2004.

Streiter O. (2001), Corpus-based parsing and treebank development, S. 115-120 in: ICCPOL 2001, 19th International Conference on Computer Processing of Oriental Languages, Seoul (Korea) 2001.

Streiter O., Voltmer L. (2003), Text Classification for Corpus-Based Legal Terminology, S. 253-260 in: Vrabie, G., Turi, J.G. (Hrsg.) (2003), La théorie et la pratique des politiques linguistiques dans le monde, Tagungsband der 8. Internationalen Konferenz der Académie Internationale de Droit Linguistique 2002, Editura CUGETAREA Iaşi (Rumänien) 2003.

Wendt H. (1987), Fischer Lexikon: Sprachen, Fischer Taschenbuchverlag, Frankfurt am Main, 1987.

Widdows D., Peters S., Cederberg S., Chan C.-K., Steffen D., Buitelaar P. (2003), Unsupervised Monolingual and Bilingual Word-Sense - Disambiguation of Medical Documents using UMLS, S. 9-16 in: Natural Language Processing in Biomedicine, ACL 2003 workshop proceedings, Association for Computational Linguistics Sapporo 2003.
http://citeseer.ist.psu.edu/584877.html : 18.6.2004.

Zitierte Rechtssprechung:
BGH-Urteil vom 11.07.2002, I ZR 255/00, zit. nach Elektronischer Pressespiegel, JurPC Web-Dok. 302/2002, http://www.jurpc.de/rechtspr/20020302.htm: 26.01.2004.

Internetquellen in Zitierreihenfolge:
http://www.eurac.edu/bistro : 1.10.2004.
http://www-opac.bib-bvb.de : 21.9.2004.
http://www.ddb.de/professionell/mab.htm : 21.9.2004
http://www.gnu.org/manual/wget/ : 23.9.2003.
http://www.w3.org/XML : 23.9.2003.
http://xml.apache.org : 23.9.2003.
http://dev.eurac.edu:8080/cgibin/bib/biblio : 18.6.2004.
http://www.cs.vassar.edu/XCES : 27.9.2003.
http://www.tei-c.org/ : 1.10.2004.
http://www.juralink.de/8LITERATUR/Umgang/Recherche.htm 27.9.2003.
http://www.bibliotek.uni-regensburg.de/rvko_neu/mytree.php3#P : 27.9.2003.
http://www.loc.gov/standards/iso639-2/normtext.html : 1.10.2004.
http://europa.eu.int/comm/eurostat/ramon/nuts/codelist_en.cfm : 27.9.2003.
http://www.gesetze.ch/ : 27.9.2003.
http://www.idg.fi.cnr.it/banche/dogi/dogi.htm : 10.10.2003.

http://www.cis.uni-muenchen.de/people/Schulz/SkriptOCR_03/OCR.pdf: 4.4.2005.
http://odur.let.rug.nl/~vannoord/TextCat/ : 10.10.2003.
http://lynx.rowser.org : 10.10.2003.

Kapitel 2

# II. Fachgebietserkennung für Terminografen

## Automatische Textklassifizierung in der korpusgestützten Rechtsterminografie

Dieses Kapitel stellt eine computerunterstützte Fachgebietserkennung für Terminografen vor. Im Theorieteil wird zunächst wird in (1.) auf die Bedeutung und Schwierigkeit der Einteilung von Terminologie in Fachgebiete eingegangen und mit der Einteilung von Texten in Fachgebiete verglichen (2.). In 3. werden der Fachgebietserkennung durch den Terminologen die computerlinguistischen Alternativen gegenübergestellt, die unter 4. aus den Erfahrungen mit den methodisch verwandten Sprachenidentifizierern beurteilt werden, insbesondere auf den Zusammenhang zwischen Erkennungsqualität und Textlänge.

Daran schließen die praktischen Versuche an. Unter Punkt 5 werden die Ergebnisse traditioneller Fachgebietszuweisung in der Rechtsdatenbank analysiert und in 6. werden die Ergebnisse einer relativ einfachen automatischen Fachgebietserkennung durch Terme vorgestellt. Unter Punkt 7 wird das Prinzip der Fachgebietserkennung von Texten durch andere Texte erläutert, das dann in 8. getestet wird. Der Punkt 9 zeigt in einem Exkurs, wie die Ergebnisse der Fachgebietserkennung zur Rekonstruktion des Zusammenhangs aller Fachgebiete benutzt werden kann. Schließlich werden die Ergebnisse evaluiert (10.) und mit einem Ausblick auf Anwendungen und Verbesserungsmöglichkeiten abgerundet (11.).

### 1. Fachgebiete von Rechtskonzepten in der Terminologie

Ein **Fachgebiet** ist der semantische Bezugsrahmen, in dem Fachbegriffe ihre spezielle Bedeutung erhalten. Die Einordnung in Fachgebiete erfolgt zweckorientiert in Anlehnung an Konventionen.[109]

Ein terminologischer Ansatz ist immer onomasiologisch (von der Bedeutung zur Lexikalisierung). Der Ansatz von BISTRO (vergl. Kapitel 1 Punkt 2) ist außerdem monodirektional (vom Italienischen ins Deutsche), mehrsprachig (Italienisch – Deutsch, Ladinisch) und mehrfach rechtsvergleichend (Rechtssysteme in Italien, Österreich, Deutschland und Schweiz).

Konkret geht man vom Rechtskonzept im italienischen Rechtssystem aus und verzeichnet seine italienischen Benennungen in der italienischen Rechtsordnung sowie entsprechende deutsche Benennungen in der italienischen, österreichischen, deutschen und schweizerischen Rechtsordnung.

Die 15.000 Rechtskonzepte gliedern sich derzeit in 24 Fachgebiete.[110] Zu Beginn der Arbeit bestimmt der Terminograf die wichtigsten Rechtskonzepte des zu bearbeitenden Fachgebiets.[111] Da-

[109] Die Einteilung von Fachbegriffen in Fachgebiete findet sich bereits 1906 in Schlohmann, A., Illustrierte Technische Wörterbücher, zit. nach S. 366-367 in: Picht H., Schmitz K.-D. (Hrsgg.) (2001), Terminologie und Wissensordnung, TermNet Wien 2001. Sie wurde 1968 durch eine ISO Norm als Standard für jede Art von Terminologie empfohlen (S. 366-367) und spielt heute eine überragende Rolle in der Terminografie, u.a. weil mit Hilfe des Fachgebiets Homonyme und Polyseme getrennt werden können (Arntz R., Picht H., Mayer F. (Hrsgg.) (2002), Einführung in die Terminologiearbeit, Studien zu Sprache und Technik Bd. 2, Georg Olms Verlag Hildesheim 2002, S. 234) und ein Überblick über die Beziehungen der Konzepte gewonnen werden kann.

durch werden die Benennungen dieses Rechtskonzepts handklassifiziert. Allerdings kann es sein, dass das Rechtskonzept nicht ausschließlich in diesem Fachgebiet verwendet wird. Es kann sogar sein, dass das Rechtskonzept in einem anderen Fachgebiet stärker im Zentrum steht.[112]

Die Einteilungen verschiedener Rechtssysteme sind nicht kompatibel. Während die Rechtswissenschaft in Deutschland und Österreich das Handelsrecht als eigenständiges Rechtsgebiet unter dem Zivilrecht und parallel zum bürgerlichen Recht auffasst, sieht die italienische Rechtswissenschaft das Handelsrecht als Teil des Zivilrechts an. Dem entspricht auch die formale Tatsache, dass in Deutschland und Österreich bereits seit dem 19. Jahrhundert eigene Handelsgesetzbücher existieren, während sich in Italien entsprechende Regelungen im *Codice Civile* (Zivilgesetzbuch) finden. Schließlich konkurrieren verschiedene Sichten des Rechts, denn nach einer Ansicht gibt es die drei Rechtsgebiete Zivilrecht, Öffentliches Recht und Strafrecht, nach einer anderen Ansicht ist das Strafrecht ein eigenständiger Teil des Öffentlichen Rechts. Schließlich konkurrieren auch Rechtstraditionen, wenn Deutschland und Österreich mit ihrem Abstraktionsprinzip und konstitutivem Grundbuch ein wichtiges Rechtsgebiet mit dem Namen Sachenrecht haben, für das es in Italien (mit Ausnahme Südtirols) und Frankreich (mit Ausnahme Elsass-Lothringens) keine Entsprechung gibt.

Wie fließend die Übergänge sind und wie diskutabel die Einteilung ist, sieht man zum Beispiel an den Konzepten „Verwaltungsstrafe“ und „Verwaltungsgebühr“. Eine Verwaltungsstrafe ist eine Strafe im Verwaltungsrecht. Eine Verwaltungsgebühr ist eine Abgabe im Verwaltungsrecht. Beide Konzepte haben jeweils ein Bein im Verwaltungsrecht und das andere in einem anderen Fachgebiet. Die „Verwaltungsstrafe“ wird mit diesem Argument zwei Fachgebieten zugeordnet (Verwaltungsrecht und Strafrecht). Andererseits wurde die „Verwaltungsgebühr“ allein dem Abgabenrecht zugeordnet (statt dem Verwaltungsrecht und dem Abgabenrecht). Argument hierfür ist, dass es praktisch in jedem Rechtsgebiet Gebühren gibt, diese Gebühren gleichzeitig aber Kern des Abgabenrechts sind. Man erkennt, dass das Argument des „Kernfachgebiets“ auch für die „Verwaltungsstrafe“ gilt und das der „doppelten Verwurzelung“ auch für die „Verwaltungsgebühr“.

Der Terminograf trifft seine Entscheidung über das Fachgebiet anhand einer Recherche in Rechtstexten, weil auch für Experten nicht sofort alle möglichen Fachgebiete eines Konzepts präsent sind. Wer etwa das Fachgebiet EU-Recht behandelt, darf nicht übersehen, dass die „Beihilfe“ auch im Verwaltungsrecht einen wichtigen Platz einnimmt und dass das Wort auch im Strafrecht ein Fachbegriff ist. Eine Hilfe bei dieser Recherche ist es, wenn die Fundstellen in Rechtstexten in Fachgebiete eingeteilt und geordnet sind.[113]

Hat der Terminograf das Fachgebiet eines zu beschreibenden Terms ausgewählt, muss er das Rechtskonzept in diesem Fachgebiet definieren. Dann muss er für alle seine Benennungen in jedem Rechtssystem einen Kontext aus zuverlässigen Quellen angeben. Sowohl Definition als auch der

---

[110] Die Rechtsgebiete sind Verwaltungsrecht, Strafrecht, Zivilrecht, Strafprozessrecht, Handelsrecht, Steuerrecht, Hochschulrecht, Zivilprozessrecht, Arbeitsrecht, Schuldrecht BT, Recht allgemein, Schuldrecht AT, Familienrecht, Straßenverkehrsrecht, öffentliches Recht, Umweltrecht, Sozialrecht, Versicherungsrecht, Gewerkschaftsrecht, Völkerrecht, Konkursrecht, Verfassungsrecht, Erbrecht und Europarecht.

[111] Zur Anfertigung von Rechtswörterbüchern siehe Bergenholtz H., Tarp S. (Hrsg.) (1995), Manual of Specialised Lexicography – The preparation of specialized dictionaries, John Benjamins Amsterdam 1995, S. 63 ff.

[112] Es sind entweder a) zwei nebeneinander liegende Fachgebiete oder b) zwei hierarchisch zueinander stehende Fachgebiete. a) Das Konzept „Straßenbenutzungsgebühr“ gehört nach dem Objekt „öffentlicher Straßenraum“ zum Straßenverkehrsrecht, von der Form „öffentliche Abgabe“ zum Steuerrecht, die beide zum öffentlichen Recht gehören. b) Das Konzept „öffentlicher Straßenraum“ ist spezieller Regelungsgegenstand des Straßenverkehrsrechts, als öffentliche Sache wird aber überall im Verwaltungsrecht darauf Bezug genommen (Widmung, Sondernutzung). Das Konzept ist also für das gesamte Verwaltungsrecht von Bedeutung, ganz besonders aber im Unterfachgebiet Straßenverkehrsrecht.

[113] Genau genommen wird hier oft das Fachgebiet der Benennung ermittelt, das nicht deckungsgleich mit dem des Konzepts sein muss, aber in der Praxis weitgehend mit ihm identifiziert wird. Siehe Ciola B. (2001), Darstellung von Äquivalenzbeziehungen in der übersetzungsorientierten Terminologiearbeit im Recht, S. 742-752 in: Mayer F. (Hrsg.) (2001), Language for Special Purposes: Perspectives for the New Millennium, Bd. 2, Narr Tübingen 2001, S. 742, der von der systematischen Beschreibung von „Begriffe[n] und Benennungen des entsprechenden Fachgebiets“ spricht.

Kontext müssen also ihrerseits möglichst im Zentrum des Fachgebiets stehen und auch dafür müssen Texte einem Fachgebiet zugeordnet werden. Die Terminografen von BISTRO konsultieren Printmedien, das lokale Korpus zwei- und dreisprachiger paralleler Gesetzestexte[114] sowie wissenschaftliche Internetquellen durch eine kontrollierte Metasuche.

### 2. Problematik der Fachgebietseinteilung von Text

Texte eines Fachgebiets enthalten oft Textteile, die für sich allein genommen einem anderen Fachgebiet zugeordnet würden. Ein gutes Beispiel ist das Nebenstrafrecht, das zum Fachgebiet „Strafrecht besonderer Teil" zu rechnen ist, sich aber in den verschiedensten Gesetzen findet. Das Lebensmittelrecht ist eine klassische Materie des besonderen Verwaltungsrechts. Gleichwohl enthält das deutsche LMBG in §§ 51 und 52 Straftaten gegen deutsches Recht und in den §§ 56 und 57 werden Verstöße gegen EU-Recht durch den deutschen Gesetzgeber mit Strafe bewehrt. Damit befinden sich Strafrechtspassagen und Europarecht im verwaltungsrechtlichen Text. Nach § 52 I Z. 10 LMBG wird bestraft, wer „mit einer irreführenden Darstellung oder Aussage wirbt". Die irreführende Werbung ist weder ein Rechtskonzept des Verwaltungsrechts noch des Strafrechts, sondern des Wettbewerbsrechts.[115] Die Benennung „irreführende Darstellung" ist in einem Abschnitt Strafrecht eingebettet, der Teil eines Lebensmittelrechtstextes ist. Diese Verschachtelung ist typisch für alle Rechtstexte. Die zunächst unerwartete Folge ist, dass die **Unterteilung eines Textes** starken Einfluss auf die Fachgebietszuordnung hat.

Viele korpuslinguistische Projekte nehmen daher nur eine formale Gliederung der Texte vor. Im Bononia Rechtskorpus wird beispielsweise eine unterschiedliche Unterteilung für italienische und englische Rechtstexte gewählt und die italienischen Rechtstexte werden nach Kodifikationen (Zivilgesetzbuch, Strafgesetzbuch, Zivilprozessbuch und Strafprozessbuch) und im Übrigen nach der Normenhierarchie geordnet.[116] Das stärker untergliederte französische Rechtskorpus der École des Mines de Paris[117] gliedert sich nur nach Kodifikationen, die dann zu Gruppen zusammengefasst sind. Solch eine formale Unterteilung hat den Vorteil, dass es nur wenige Klassen gibt, und dass die Klassen klar abgegrenzt und verständlich sind.

Eine Unterteilung nach Kodifikationen kann jedoch nur wenige Hierarchiestufen abbilden. Zivilgesetzbücher enthalten fast immer große und wichtige Untergebiete wie Familienrecht und Erbrecht. Wie alle Fachgebiete haben diese jedoch eine eigene Terminologie, deren Bedeutung sich von Kapitel zu Kapitel eines Gesetzbuchs ändern kann. In den genannten Rechtskorpora wird dem nicht Rechnung getragen.

Ein weiterer Nachteil ist, dass die formale Beschreibung zu anderen Ergebnissen führt als die inhaltliche Beschreibung. Beispiel hierfür ist wieder das Nebenstrafrecht. Die §§ 51, 52 LMBG sind formal Lebensmittelrecht, inhaltlich aber Strafrecht. In der Terminografie kommt es im Gegensatz zur Lexikographie gerade auf die inhaltliche Beschreibung der Konzepte an. Es wäre ein klarer Kunstfehler, wenn zur Beschreibung des Begriffs „Lebensmittel" eine (neben-) strafrechtliche Bestimmung zitiert würde.

Texte enthalten ebenso wie Begriffe, und meist sogar noch stärker, die Aspekte von verschiedenen, oft ineinander verschachtelten Fachgebieten. Die Zuordnung von Texten zu Fachgebieten, wie

---

[114] Gamper J. (1999), Construction of a Parallel Text Corpus Encoding Primary Data, in: Academia Nr: 18, (März - Juni 1999), EURAC, Bozen 1999, http://www.eurac.edu/Press/Academia/18/Art_13.asp :30.3.2004.

[115] Formal in § 3 des deutschen Gesetzes über den unlauteren Wettbewerb (UWG) geregelt. Inhaltlich spielt wegen der Richtlinie 84/450/EWG des Rates vom 10.9.1984 zur Angleichung der Rechts- und Verwaltungsvorschriften der Mitgliedstaaten über irreführende Werbung aber das Europarecht die tragende Rolle.

[116] Bononia Legal Corpus (BoLC) http://www.cilta.unibo.it/Portale/RicercaLinguistica/bolc_eng.html : 1.10.2004.

[117] Centre de Recherche en Informatique, École des Mines de Paris
http://www.cri.ensmp.fr/info_juridique/projet_info_juridique.html : 1.10.2004, http://ontologie.w3sites.net/ : 1.10.2004.

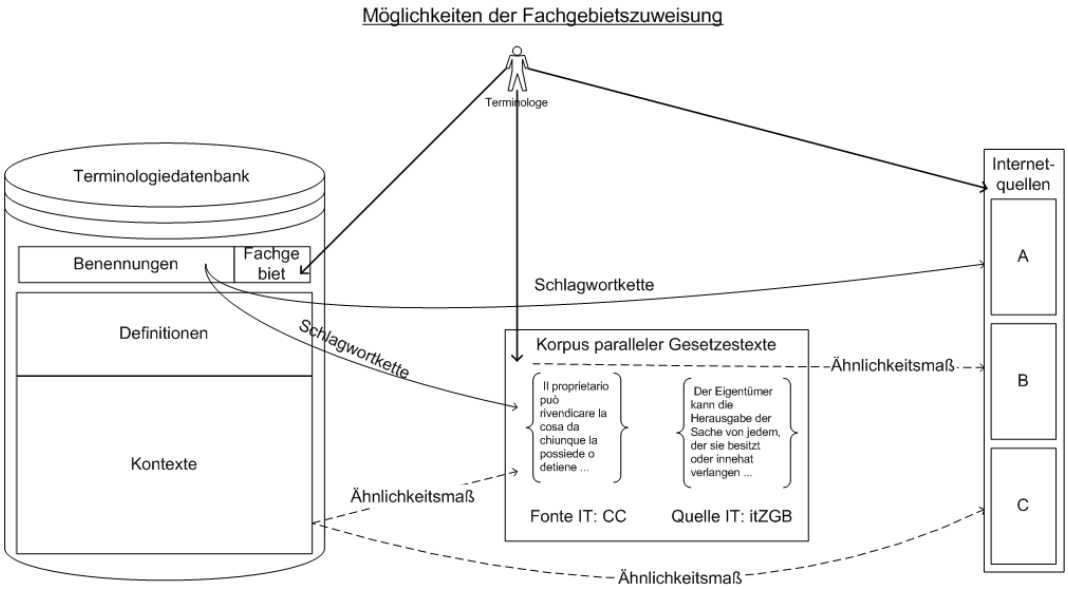
sie etwa für die Korpuserstellung notwendig ist, bereitet daher noch größere Schwierigkeiten als die Zuordnung von Rechtskonzepten zu Fachgebieten.

### 3. Von intellektueller zu automatischer Fachgebietseinteilung

Traditionell bleibt es dem Terminologen überlassen, das Fachgebiet der Konzepte und der Texte aus dem Korpus, den Print- oder Internetquellen herauszufinden. Eine nach Fachgebieten geordnete Präsentation der Texte könnte den Terminografen bei der Suche nach geeigneten Texten zur Beschreibung des Konzepts unterstützen.

Printmedien sind zwar geordnet, das Finden von Kontexten ist aber sehr zeitraubend. Ein elektronisches Korpus erleichtert das Finden von Kontexten, ihre Strukturierung nach Fachgebieten setzt aber voraus, dass die Texte durch den Terminografen oder geeignete Verfahren zugeordnet werden. In Internetquellen sind Kontexte noch leichter zu finden, sie müssen aber außerdem auf ihre Seriosität und Fachlichkeit geprüft werden und die nötige Zuordnung kann nicht gespeichert werden wie beim Korpus, sondern muss bei jeder Quelle erneut spontan erfolgen.

**Grafik 7: Möglichkeiten der Fachgebietserkennung**




Das Schaubild 7 zeigt oben die intellektuelle Fachgebietseinordnung durch den Terminologen. Er weist den Benennungen in der Terminologiedatenbank ein Fachgebiet zu und sucht fachspezifische Definitionen und Kontexte zu dieser Benennung. Der Terminologe findet die Benennung im Korpus (Mitte) oder in Internetquellen (rechts) und muss entscheiden, ob es sich um Fachtexte im gesuchten Fachgebiet handelt.

Entlang der gestrichelten Linien werden klassifizierte Texte mit Texten ohne Fachgebietsinformation verglichen. Die Hypothese ist, dass die Fließtextstücke natürlicher Sprache, deren Zeichen sich am stärksten ähnlich sind, auch dem gleichen Fachgebiet angehören,[118] so dass die Fachgebietsinformation in Pfeilrichtung übertragen wird. Solch ein Textvergleich funktioniert zwischen fachgebietsspezifischen Kontexten und Korpustexten, zwischen fachgebietsspezifischen Kontexten und Internettexten sowie zwischen klassifizierten Korpustexten und Internettexten.[119] Um ausreichend

[118] Zur automatischen Klassifikation siehe S. 2.

[119] Wie der Anfangspunkt der gestrichelten Linie andeutet, bietet sich in erster Linie der Vergleich von handklassifizierten Korpustexten mit Internetquellen an. Wenn das Korpus bereits mit statistischen Mitteln klassifiziert wird, könnten sich durch zweimalige Anwendung eines Ähnlichkeitsmaßes Fehler anhäufen bzw. es könnte Ähnlichkeitsinformation durch den Zwischenschritt verloren gehen.

Text zu haben, können Kontexte der Datenbank und Korpustexte zu einem Pseudokorpus zusammengefasst werden, um Internetquellen zu identifizieren.

Zwischen der traditionellen Fachgebietszuweisung und der Text zu Text-Methode gibt es die dritte Möglichkeit, von Termen auf Text zu schließen. Die Hypothese ist, dass das häufige Auftreten von Benennungen eines Fachgebiets in einem Textstück darauf hindeutet, dass auch das Textstück aus diesem Fachgebiet ist. Dazu müssen die Datenbankterme als Schlagwörter indexiert und das zu klassifizierende Textstück mit diesem Index verglichen werden. Die Texte werden wie eine Anhäufung bereits bekannter Fachbegriffe behandelt und das vorherrschende Fachgebiet der Benennungen wird als Fachgebiet des Textes vermutet. Die Benennungen können als Schlagwortkette verstanden werden, weil nicht eine einzelne Benennung, sondern nur das Zusammenspiel mehrerer Benennungen zu einem zuverlässigen Ergebnis führt.[120] Eine Voraussetzung dieses Ansatzes sind viele von Terminografen hergestellte Beziehungen Benennung - Fachgebiet, damit in jedem beliebigen Fachtext genügend Benennungen vorkommen.[121]

Beide automatischen Ansätze sind statistische, also nicht regelbasierte Ansätze.[122] Damit muss die in der Fachsprachenforschung so umstrittene Frage, was einen Text als Fachtext in einem bestimmten Gebiet auszeichnet, nicht explizit beantwortet werden.[123] Das Ergebnis der Verfahren ist die Übertragung bereits angesammelten Wissens über die Fachsprachlichkeit von Texten bei der Text-Text Methode und von Begriffen bei der Term-Text Methode.

### 4. Fachgebietserkennung und Sprachenidentifizierung

Je mehr Benennungen im zu klassifizierenden Text gefunden werden, umso mehr Information steht der Fachgebietserkennung zur Verfügung und umso eher kann man gute Ergebnisse erwarten. Die Fachgebietserkennung ist also direkt abhängig von der Anzahl der Benennungen und der Textlänge.

Bei dem verwandten Problem der Sprachenidentifizierung gilt die gleiche Abhängigkeit. Sprachenidentifizierer arbeiten entweder mit *n*-Grammen oder mit Wörterbüchern. Beim Wörterbuchansatz werden die Benennungen im Wörterbuch indexiert. Man vergleicht dann die Wörter des Eingabetextes mit den Indexen für die verschiedenen Sprachen und die größte Übereinstimmung gibt die Sprache an.

Die Sprachenidentifizierung von einsprachigen Texten mit mindestens zwanzig Wörtern natürlichen Textflusses ohne zu viele Eigennamen, Abkürzungen u.ä. gilt seit Jahren als gelöstes Problem.[124] Es gibt aber zwei wesentliche Unterschiede zur Fachgebietserkennung.

Erstens ist der Eingabetext zur Fachgebietserkennung im Unterschied zur Sprachenidentifizierung nicht ‚einsprachig' in dem Sinn, dass er nur aus Bausteinen eines Wörterbuchs zusammenge-

---

[120] Schlagwortketten und Indexe werden auch beim Dokumentretrieval benutzt. Dort kommt es jedoch auf die diskriminierende Funktion von Wörtern an. Ziel beim Retrieval ist es, die Charakteristik von Texten im Verhältnis zu anderen Texten herauszufinden (relativ), bei der Fachgebietszuweisung wird hingegen die Fachlichkeit eines Textes unabhängig von der Fachlichkeit oder Menge anderer Texte (absolut) gesucht.

[121] Die Datenbank BISTRO enthält derzeit ca. 65.000 Beziehungen zwischen Benennungen und Fachgebiet.

[122] Eine der ersten Ansätze zum Einsatz von statistischen Verfahren in der Fachsprachenforschung ist Hoffmann L. (1975), Fachsprachen und Sprachstatistik. Beiträge zur angewandten Sprachwissenschaft, Akademieverlag Berlin 1975.

[123] Das Problem der Abgrenzung und Definition von Fachsprachen ist seit längerem bekannt, aber bisher ungelöst. Hierzu Hahn W. v. (1983), Fachkommunikation: Entwicklung, linguistische Konzepte, betriebliche Beispiele, Sammlung Göschen 2223, de Gruyter Berlin, New York 1983, S. 60 ff.

[124] Vergl. etwa die Ergebnisse von Grefenstette G. (1995), Comparing Two Language Identification Schemes, S. 263-268 in: Bolasco S., Lebart L., Salem A. (Hrsgg.) (1995), Proceedings of the 3. International Conference on Statistical Analysis of Textual Data (Journées d'Analyse de Données Textuelles JADT 95), CISU Rom 1995, http://www.xrce.xerox.com/Publications/Attachments/1995-012/Gref---Comparing-two-language-identification-schemes.pdf : 8.4.2004. Der Sprachenidentifizierer liegt unter http://www.xrce.xerox.com/competencies/content-analysis/tools/guesser : 8.4.2004.

setzt wäre. Die meisten Rechtstexte enthalten nicht nur Fachbegriffe eines Fachgebiets, sondern benutzen auch die ‚Sprache' anderer Fachgebiete zur Kommunikation. Zur Bestimmung des Rechtswegs in einer Prozessordnung muss zwangsläufig das materielle Recht genannt werden, auf das sich die Rechtswegsbestimmung beziehen soll.[125] Die Ausgangssituation ist also eher mit der Erkennung der vorherrschenden Sprache in einem vielsprachigen Text zu vergleichen. Allerdings müssen Sprachenidentifizierer bis zu hundert Sprachen erkennen, der Fachgebietserkenner in unserem Fall nur 24.

Zweitens sind bei der Sprachenidentifizierung die hochfrequenten Funktionswörter ausschlaggebend, während die Fachgebietserkennung die allgemein häufigen Wörter gerade ausschließt und sich auf Fachbegriffe mittlerer und niedriger Frequenz stützt. Im Unterschied zur Fachgebietserkennung werden Sprachenidentifizierungswörterbücher daher aus Fließtext hergestellt, damit es die besonders häufigen Funktionswörter und Wortformen enthält. Es genügen dann auch nur wenige tausend Wörter pro Sprache als Lernmaterial, um die meisten Wörter im entsprechenden Wörterbuch wiederzufinden.[126] Dieser Zusammenhang zwischen Textmenge und Klassifizierungsqualität wird vom Zipf'schen Gesetz bestätigt und ist daher gut übertragbar.[127] Es ist also zu erwarten, dass eine Erhöhung der Benennungen nur schwach zu neuen Treffern beiträgt. Zwar ist auch die Erhöhung der Textmenge dieser Gesetzmäßigkeit unterworfen, hier zählen aber nicht nur neue Treffer, sondern auch wiederholte Treffer, so dass ein zigmal gebrauchter Fachbegriff auch zigfach in die Fachgebietsbestimmung eingeht. Es ist also zu erwarten, dass die Qualität der Klassifikation mit der Länge des Eingabetextes wächst.

Ein Sprachenidentifizierer kann mit etwa zwölf erkannten Wörtern eine fast immer richtige Entscheidung zwischen 100 Sprachen fällen. Entsprechend könnte man für die Fachgebietserkennung Texte fordern, in denen zwölf Fachbegriffe eines Fachgebiets vorkommen. Solche Texte müssten einige Seiten Länge haben und für kurze Belegtexte, wie sie in der Terminografie verwendet werden, könnte dann grundsätzlich keine Fachgebietserkennung stattfinden. Daher sollen hier auch kurze Texte zugelassen sein, wobei man in Kauf nehmen muss, dass die Qualität mit der Textlänge abnimmt.

In BISTRO können Terme im Korpus oder auf ausgewählten Internetseiten gesucht werden. Das Korpus ist in einzelne Sätze oder kurze Paragrafen zerlegt, die selten für eine sichere Fachgebietserkennung ausreichen, die durchschnittliche Internetseite bietet schon bessere Aussichten. Als dritte Möglichkeit bietet BISTRO die Eingabe eines freien, auch selbst geschriebenen Textes beliebiger Größe an, über den man dann die Fachgebietserkennung laufen lassen kann.

### 5. Ausgangsdaten zur Fachgebietserkennung

Wie bei der Sprachenidentifizierung ist auch bei der Fachgebietserkennung quantitativ und qualitativ gutes Ausgangsmaterial erforderlich. Die nötige Quantität bemisst sich dabei nach der Anzahl der Fachgebiete, die erkannt werden soll. Es muss also in jedem einzelnen Fachgebiet hinreichend Lernmaterial vorhanden sein, damit die Unterschiede zu den anderen Fachgebieten charakteristisch hervortreten können. Die Qualität des Lernmaterials betrifft die oben diskutierte unsichere Einteilung in Fachgebiete, sobald ein Konzept mehr als eine Affinität ausgebildet hat. In der Statistik schlägt sich dies in häufigerer Zuteilung zu mehreren Fachgebieten nieder. Eine Analyse der Ausgangsdaten von BISTRO ergab die in Grafik 8 gezeigte Zusammensetzung:

---

[125] § 23 deutsches GVG: „Die Zuständigkeit der Amtsgerichte umfasst […] Ansprüche aus einem Mietverhältnis über Wohnraum, […] über Wirtszechen, Fuhrlohn, Überfahrtsgelder […]; das Aufgebotsverfahren." Das einzige Wort des Prozessrechts ist „Zuständigkeit", während dutzende Begriffe des materiellen Zivilrechts vorkommen.

[126] Langer (2002) a.a.O., S. 103.

[127] Nach dem Zipf'schen Gesetz gibt es wenig häufige und sehr viele seltene Benennungen.

**Grafik 8: Anteil eindeutiger Benennungen nach Fachgebiet**

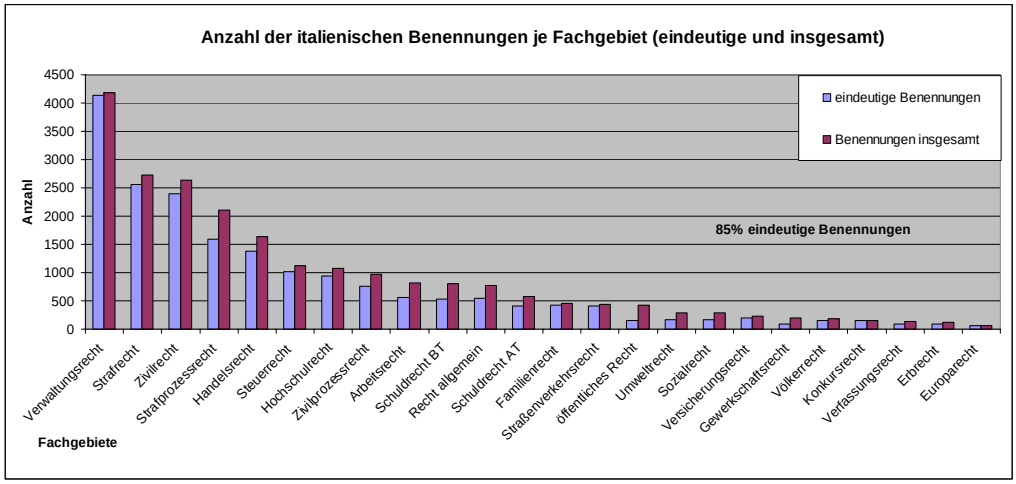



**Tabelle 5: Legende zu den Fachgebietsbezeichnungen**

| | | |
|---|---|---|
| 1. | juris | Recht allgemein |
| 2. | civile | Zivilrecht |
| 3. | pubblico | Öffentliches Recht |
| 4. | lavoro | Arbeitsrecht |
| 5. | commerciale | Handelsrecht |
| 6. | tributario | Abgabenrecht |
| 7. | penale | Strafrecht |
| 8. | processuale penale | Strafprozessrecht |
| 9. | processuale civile | Zivilprozessrecht |
| 10. | sociale | Sozialrecht |
| 11. | universitario | Hochschulrecht |
| 12. | costituzionale | Verfassungsrecht |

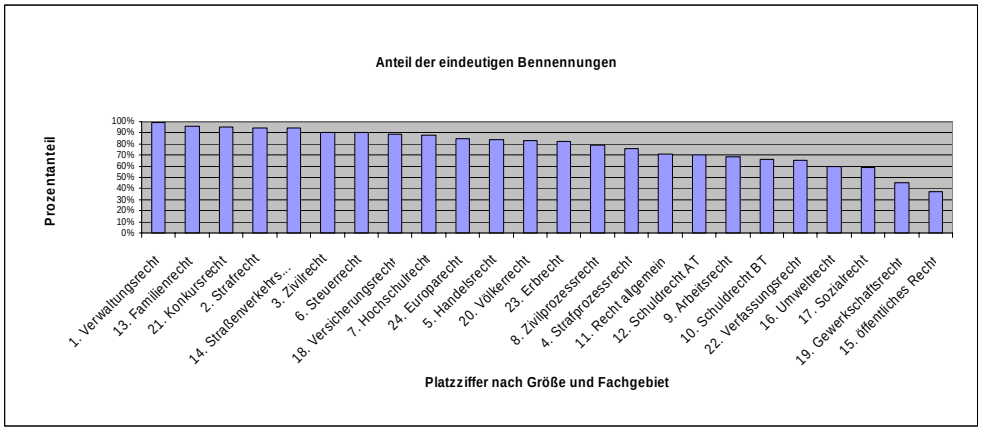
| | | |
|---|---|---|
| 13. | ambiente | Umweltrecht |
| 14. | obbligazioni parte speciale | Schuldrecht allg. Teil |
| 15. | comunitario | EU-Recht |
| 16. | obbligazioni | Schuldrecht allg. Teil |
| 17. | internazionale | Völkerrecht |
| 18. | famiglia | Familienrecht |
| 19. | sindacale | Gewerkschaftsrecht |
| 20. | navigazione | Seefahrtsrecht |
| 21. | assicurazioni | Versicherungsrecht |
| 22. | stradale | Straßenverkehrsrecht |
| 23. | fallimentare | Insolvenzrecht |
| 24. | successioni | Erbrecht |

Die Benennungen[128] sind sehr ungleich auf die Fachgebiete verteilt. Das größte Fachgebiet hat über 4000 italienische Benennungen, das kleinste weniger als Hundert. Trotz der oben erörterten Schwierigkeit der exklusiven Zuordnung von Rechtskonzepten gehört im Durchschnitt nur jede sechste Benennung mehr als einem Fachgebiet an. Das ist im Verhältnis zu der Anzahl der Fachgebiete wenig, denn die Information, dass eine Benennung diese zwei Fachgebiete hat, nicht aber eines der 22 anderen ist immer noch eine starke Aussage.

Als nächstes wurde untersucht, ob eher große oder kleine Fachgebiete besonders viele mehrdeutige Benennungen aufweisen. Wenn man die geographische Metapher von ‚Fachgebiet' ernst nimmt, dann hätten größere Gebiete im Verhältnis zu ihrem Inhalt kürzere Grenzen als kleine Gebiete und könnten daher mehr eindeutige Benennungen enthalten. Tatsächlich ergibt sich folgendes Bild:

[128] Die Fachgebietseinteilung erfolgt nach den italienischen Rechtskonzepten, denen dann rechtsvergleichend die deutschen Termini zugeordnet werden. Es kann also vorkommen, dass die entsprechenden deutschsprachigen Termini einem anderen Fachgebiet angehören. Um diese Folge des monodirektionalen Ansatzes auszuschließen, werden unten in den Versuchen nur die italienischen Daten verwendet und Fachgebiete des italienischen Rechtssystems zu erkennen versucht. In Tabellen und Grafiken wird zum leichteren Verständnis vorzugsweise die deutsche Fachgebietsbezeichnung verwendet.

**Grafik 9: Anteil eindeutiger Benennungen je Fachgebiet**




Die Korrelation zwischen zwei Rangreihen, hier also der Mehrdeutigkeit und Größe von Fachgebieten, kann mit der Spearman-Korrelation gemessen werden.[129] Sie kann Werte zwischen 1 (völlige Übereinstimmung) und -1 (perfekter negativer Zusammenhang) annehmen. Hier ist der Wert 0,322, so dass bei der kurzen Datenreihe kein Zusammenhang nachgewiesen ist.

Eine Erklärung für die ungleiche Verteilung mehrdeutiger Rechtskonzepte auf Fachgebiete wäre, dass einige Fachgebiete so intensiv behandelt wurden, dass auch marginale, zweideutige Konzepte Aufnahme fanden, während bei anderen Fachgebieten nur die Kernkonzepte aufgenommen wurden. Eine andere Erklärung könnte die unterschiedliche Auslegung der Klassifikationsrichtlinien sein (s.o.).

### 6. Direkte Textfachgebietserkennung durch Termini

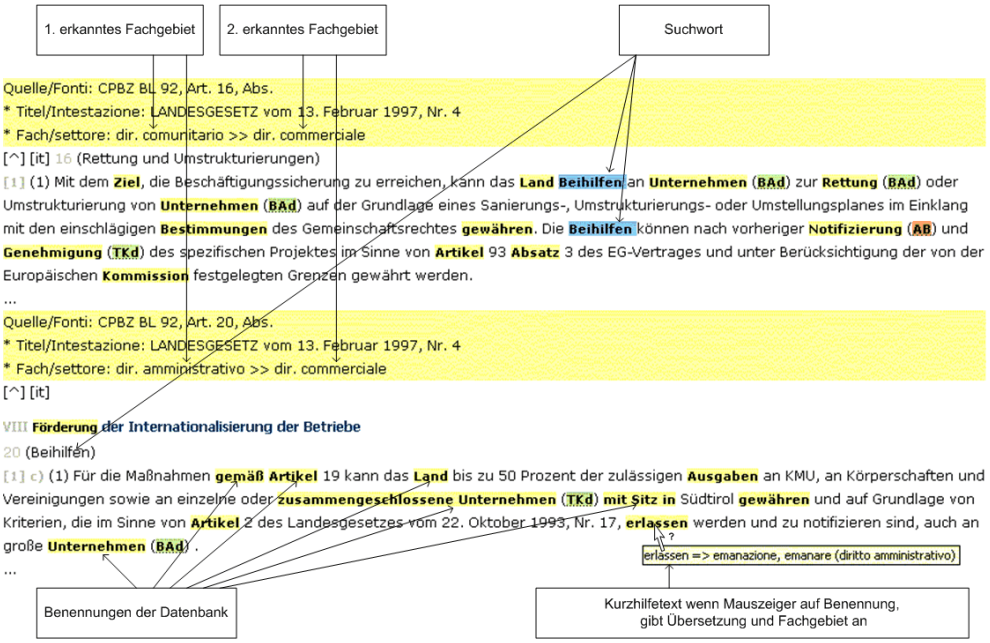
Um die intellektuelle, originäre Zuweisung eines Fachgebiets durch eine automatische Methode zur Erkennung des Rechtsgebiets eines Fachtextes zu ersetzen, können Fachbegriffe direkt in den Texten gesucht werden. Dazu wird der einzuordnende Textabschnitt mit dem Index aller Datenbanktermini verglichen. Jeder Term der im Text vorkommt verleiht dem Text ein gewisses ‚Fachgewicht'. Hat ein Term mehrere Fachgebiete, kann er mehrere Fachgewichte zuweisen, allerdings nur anteilig, denn sonst würden diese Benennungen stärker wiegen als Benennungen mit einem Fachgebiet. Auch der Effekt, dass Fachbegriffe aus großen Fachgebieten häufiger gefunden werden, wird durch anteilige Gewichtung ausgeglichen.[130] Die Fachgewichte werden aufsummiert und der höchste Wert bestimmt das Fachgebiet. Ist die Differenz zwischen den beiden höchsten Werten sehr klein, dann kennzeichnen beide Fachgebiete den Text. Die Häufigkeit von Fachbegriffen im vorgelegten Text stuft ihn zugleich als Rechtstext oder aber als anderer Fach- oder Allgemeintext ein. Wird ein festzulegender Schwellenwert nicht erreicht, dann wird auch kein Fachgebiet zugewiesen. Es folgen zwei Beispiele für die Fachgebietserkennung durch Termini.

Im ersten Beispiel wird der Begriff „Beihilfe" gesucht, der in strafrechtlichen Texten als untergeordneter Beitrag zu einer fremden Straftat auftauchen kann, oder im Verwaltungs- und EU-Recht als von der Verwaltung vergebener finanzieller Vorteil. Die Korpussuche in der Südtiroler Landesgesetzgebung bringt Art. 16 und 20 eines Landesgesetzes von 1997 auf den Bildschirm (Grafik 10).

[129] „Der Zusammenhang zweier ordinalskalierter Merkmale wird durch die Rangkorrelation nach Spearman ($r_s$ oder rho) erfasst." Bortz J. (1993), Statistik für Sozialwissenschaftler, 4. Auflage Springer Berlin u.a. 1993, S. 214.

[130] Die Hälfte aller Begriffe gehört dem Recht allgemein an, je ein Sechstel sind Zivil- und Strafrecht. Eine andere Möglichkeit wäre der Vergleich der Begriffe mit einem klassifizierten Hintergrundkorpus.

**Grafik 10: Beispiel der Fachgebietserkennung im Korpus, hier Suche in der Südtiroler Landesgesetzgebung nach ‚Beihilfe'**

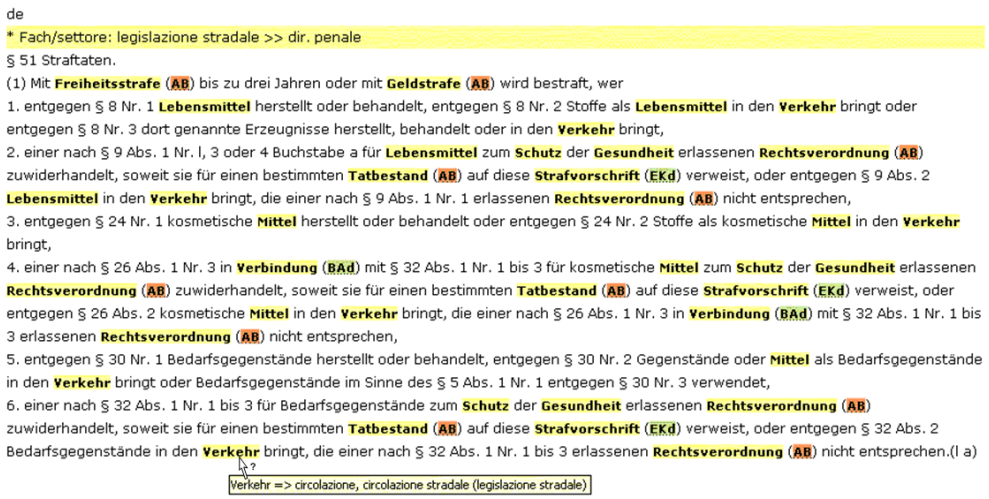



Der erste Textabschnitt benutzt das Wort „Beihilfen" im verwaltungsrechtlich-europarechtlichen Sinn. Inhalt und Ausmaß der Beihilfen wird vom Europarecht bestimmt, formal werden sie von der Verwaltung vergeben. Über die Termini wurde der Text richtig als EU-Recht (*dir. comunitario*) eingestuft. Als zweitwahrscheinlichstes Fachgebiet wird Handelsrecht (*dir. commerciale*) genannt, weil der Term „Unternehmen" als handelsrechtlicher Fachbegriff in der Datenbank ist. Dieses Fachgebiet ist falsch, der Fehler ist aber nachvollziehbar.

Der zweite Textabschnitt desselben Textes wird als verwaltungsrechtlich und ansonsten handelsrechtlich eingestuft. Das erste Fachgebiet ist richtig erkannt, beim zweiten hat wieder das Wort „Unternehmen" irregeführt.

An diesem Beispiel kann gezeigt werden, dass die Methode grundsätzlich funktioniert.

**Grafik 11: Beispiel der Fachgebietserkennung einer Internetquelle, hier § 51 LMBG**




Im zweiten Beispiel soll demonstriert werden, warum die Fachgebietserkennung durch Termini fehleranfällig ist. Der in Grafik 11 gezeigte Text gehört zum so genannten Nebenstrafrecht, also zu Strafbestimmungen, die nicht ins Strafgesetzbuch aufgenommen wurden. Das Lebensmittel- und

Bedarfsgegenständegesetz (LMGB) insgesamt gehört zum besonderen Verwaltungsrecht, sein § 51 enthält aber Strafrecht, so dass die formale Einteilung (vergl. bei Fußnote 117) nach Kodifikation scheitert.

Aufgrund der Termini wird der Textabschnitt überraschenderweise als Straßenverkehrsrecht eingeschätzt, weil die sieben Mal auftauchende Benennung „Verkehr" nur dem Fachgebiet Straßenverkehrsrecht zugeordnet ist. Zwar enthält die Datenbank die strafrechtlichen Benennungen „Inverkehrbringen"[131] und „Falschgeld in Verkehr bringen", diese wurden hier jedoch nicht erkannt. Obwohl die Benennungen „Freiheitsstrafe", „Geldstrafe", „Lebensmittel", „Tatbestand", „Strafvorschrift" und sogar das Wort „Mittel"[132] (hier fälschlicherweise, weil es um kosmetische Mittel geht!) ein strafrechtliches Fachgewicht einbringen, wird der Text als Straßenverkehrsrecht eingestuft. Grund ist das fünfmal höhere Gewicht, das jeder Term aus den knapp 500 straßenverkehrsrechtlichen Benennungen einbringt im Verhältnis zu einem der über 2500 strafrechtlichen Begriffe.[133] Das zweitwahrscheinlichste Fachgebiet ist dann das richtige Fachgebiet Strafrecht. Auf das formal nahe liegende Verwaltungsrecht deuten keine Begriffe hin.

Das zweite Beispiel zeigt, dass die Methode sich von Formalien abwendet und sich ganz auf die in der Datenbank gespeicherten Bedeutungen der Termini verlässt. Dadurch wird das Verfahren stark fehleranfällig, wenn Wörter mehrere Bedeutungen haben (Polyseme), von denen nicht alle in der Datenbank verzeichnet sind. Es ist insbesondere schädlich, wenn ein Term im konkreten Kontext nicht als Rechtsbegriff verwendet wird (Mittel nicht als Tatmittel sondern als kosmetisches Mittel), oder wenn ein Rechtsbegriff in einer Weise verwendet wird, dass er nicht als solcher erkannt wird („in den Verkehr bringen" wird nicht als „Inverkehrbringen" erkannt).

Die wissenschaftliche Suchmaschine Scirus[134] klassifiziert ebenfalls Fachtexte mit der Wortmethode. Die Termdatenbank wird durch korpuslinguistische und intellektuelle Methoden aufgebaut und in 20 Fachgebiete[135] eingeteilt. Aus diesem Korpus wissenschaftlicher Texte werden nach Klassifizierungskraft und anderen Kriterien Ein- und Mehrwortterme extrahiert. Zusätzlich wird das Wörterbuch der einzelnen Fachgebiete mit Benennungen aus Fachwörterbüchern angereichert und manuell bearbeitet.[136]

---

[131] Inverkehrbringen ist zwar dem Duden von 2003 mit der neuen Rechtschreibung nicht bekannt, die substantivierte Form findet sich aber weiterhin als Schlüsselbegriff im deutschen und österreichischen Lebensmittelgesetz, vergl. die Legaldefinitionen in § 7, Abs. 1 Spiegelstrich 2 deutsches LMBG und § 1, Abs. 2 österreichisches Lebensmittelgesetz (LMG) (dort in Abs. 1 „In-Verkehr-Bringen" geschrieben. In der Schweiz wird der Begriff Inverkehrbringen zwar nicht im Lebensmittelgesetz, aber an verschiedenen anderen Stellen legaldefiniert und tausendfach in Normen verwendet: http://search.admin.ch/cgi-bin/query?mss=de%2Fsimple&pg=q&what=web&enc=iso88592&fmt=.&q=Inverkehrbringen&submit=Suche: 8.9.2004.

[132] Ein Mittel oder Tatmittel im strafrechtlichen Sinn sind die für die Tat bestimmten Dinge.

[133] Zur Verbesserung der Ergebnisse müssten Fachbegriffe, die zugleich Allgemeinbegriffe sind, aus dem Wörterbuch ausgeschlossen werden. Zwar wirkt sich der Effekt, dass Fachbegriffe zugleich Allgemeinbegriffe sind, auf alle Fachgebiete in gleicher Weise aus, aber im Einzelfall verzerren sie die Ergebnisse doch erheblich.

[134] http://www.scirus.com/ : 21.1.2004.

[135] Agricultural and Biological Sciences, Astronomy, Chemistry and Chemical Engineering, Computer Science, Earth and Planetary Sciences, Economics, Business and Management, Engineering, Energy and Technology, Environmental Sciences, Languages and Linguistics, Law, Life Sciences, Materials Science, Mathematics, Medicine, Neuroscience, Pharmacology, Physics, Psychology, Social and Behavioral Sciences, Sociology.

[136] "Scirus maintains a customized linguistic knowledge base for each subject area. The vocabulary on the pages is mapped against the terms from dictionaries which have been compiled through training on a very large, manually pre-classified corpus of scientific texts. The dictionaries are supplemented with entries from domain-specific terminological databases. […] The vector terms are weighted single word and multiword expressions. The weight of the classification terms is determined by examining statistical properties from the training corpus – such as the classification strength – and through partial manual maintenance." http://www.scirus.com/press/pdf/WhitePaper_Scirus.pdf: 30.1.2004, S. 10.

Es wird also von Trainingskorpora auf das Fachgebiet der Terme geschlossen und dann von den Termen auf das Fachgebiet unbekannter Texte. Dieser zweite Schritt gleicht der eben beschriebenen Aufgabe und Lösungsmethode. Allerdings unterscheiden sich die Arbeitssprachen und die verwendeten Klassen, so dass es keine Schnittmenge an Dokumenten gibt, für die man Versuchsergebnisse von Scirus und BISTRO vergleichen könnte.

Scirus vergleicht die 20 Indizes mit den zu klassifizierenden Texten, wobei Treffern im Titel oder der URL besonderes Gewicht verliehen wird. Auch Metainformationen wie der Beschreibungstext hinverweisender Links fließen in die Klassifizierung ein.[137] Die Indexwörterbücher werden außerdem normalisiert und dienen zugleich als Schlagworte, mit denen die Texte zu Retrievalzwecken inhaltlich beschrieben werden.

Allerdings werden nicht alle Seiten einem der Fachgebiete zugeordnet. Der Suchbegriff „Rechtsterminologie“ fand über dreihundert Dokumente, die gleiche Suche in den 20 Fachbereichen nur etwas über 50 Dokumente. Die Wörterbücher sind stark auf englischsprachige Texte ausgerichtet, die häufiger zugeordnet werden können. Auch bei dem Begriff *penal law* (Strafrecht) werden von 9.864 Texten nur 7.250 zugeordnet und davon nur 5.369 dem Fachgebiet *law* (Recht). Die Gründe für das Nichtklassifizieren dürften darin zu suchen sein, dass die zu klassifizierenden Texte und die bereits klassifizierten zu wenig gemeinsame Wörter aufweisen. Das spricht dafür, eine andere Methode zu verwenden, um auch kleinere Übereinstimmungen zu finden und die Quantität der Daten für die Gleichheitsberechnung zu erhöhen. Ein solches Verfahren ist die in Kapitel 1 beschriebene *n*-Gramm-Methode.

## 7. Vorversuche zur Fachgebietserkennung durch Textvergleich

### 7.1. Vorüberlegungen zur statistischen Aussagekraft und Validität

Wie aus den obigen Beschreibungen der Fachgebietserkennung durch klassifizierte Termini deutlich wird, sind Termini im Prinzip im Stande, das Fachgebiet eines Textes richtig zu erkennen. Da dieser Ansatz aber auf statistischen Berechnungen beruht, müssen deren Grundannahmen erfüllt sein, was umso kritischer wird, je weniger Daten vorhanden sind, also je kürzer ein Text ist. Wenn ein Text kein Fachwort enthält, das auch in der Datenbank ist, dann kann er mit der oben beschriebenen Methode auch nicht klassifiziert werden.

Derselbe Text kann aber möglicherweise klassifiziert werden, wenn man nicht von Termini, sondern von Texten ausgeht, denn aufgrund der größeren Anzahl Wörter ist es sehr viel wahrscheinlicher, dass ein Wort beiden Texten angehört. In der Praxis wird es sogar sehr viele Wörter geben, die beiden Texten angehören, so dass man wie bei obiger Methode aus der Frequenz Schlüsse ziehen kann. Dazu erstellt man die Frequenztabelle zum klassifizierten Text und vergleicht sie mit der Frequenztabelle des unbekannten Textes wie in Tabelle 6.

[137] Zur Klassifizierungskraft von Links und Metainformationen siehe Streiter O., Voltmer L. (2003), Text Classification for Corpus-Based Legal Terminology, S. 253-260 in: Vrabie G., Turi J. G. (Hrsgg.) (2003), La théorie et la pratique des politiques linguistiques dans le monde, Tagungsband der 8. Internationalen Konferenz der Académie Internationale de Droit Linguistique 2002, S. 253-260, Iaşi (Rumänien).

**Tabelle 6: Frequenztabellen zweier Texte**

| $Wort_1$ $Wort_2$ $Wort_3$ $Wort_4$ $Wort_5$ $Wort_6$ ... | |
|---|---|
| Wort: | Frequenz: |
| $Wort_1$ | 14 |
| $Wort_2$ | 1 |
| $Wort_3$ | 1 |
| $Wort_4$ | 3 |
| ... | ... |

| $Wort_2$ $Wort_9$ $Wort_{13}$ $Wort_{14}$ $Wort_2$ … | |
|---|---|
| Wort: | Frequenz: |
| $Wort_2$ | 11 |
| $Wort_9$ | 1 |
| $Wort_{13}$ | 1 |
| $Wort_{14}$ | 34 |
| ... | ... |

Man kann solche Frequenztabellen nicht nur für die Wörter der Texte erstellen, sondern auch für größere Zeichenabschnitte des Textes wie Termini, oder kleinere Zeichenabschnitte wie *n*-Gramme.

Eine Voraussetzung für eine korrekte Fachgebietserkennung ist, dass ähnliche Texte eine statistisch aussagekräftige Ähnlichkeit in der Frequenzverteilung aufweisen. Die im Text verwendeten Zeichenfolgen (Termini, Wörter, *n*-Gramme) müssen also repräsentativ sein.[138] Je kürzer der zu klassifizierende Text ist, umso weniger Merkmale enthält er. Bleibt man bei Termfrequenzen, dann kann es passieren, dass keine Übereinstimmung vorliegt. Selbst übereinstimmende Wortfrequenzen können so niedrig sein, dass das Ergebnis völlig verzerrt wird. Der Frequenzvergleich ist keine qualitative Methode mehr, sondern eine quantitative, die umso besser funktioniert, je mehr Daten verglichen werden können. Die Berechnung der Ähnlichkeit von zwei Texten beruht also auf der statistischen Repräsentativität der verwendeten Zeichenketten.

Je länger die *n*-Gramme, umso eher werden sie valide Indikatoren, die zahlenmäßig unzureichend sind. Die Frequenztabelle wird sehr flach sein, also viele verschiedene *n*-Gramme mit niedriger Frequenz enthalten. Je kürzer die *n*-Gramme gewählt werden, umso stärker differenzieren sie sich durch ihre Frequenz, was eine statistische Gewichtung und damit ein repräsentatives Bild ermöglicht. Man kann so die repräsentativen *n*-Gramme herausfinden, büßt tendenziell aber an Validität ein.

### 7.2. Experiment zur Fachgebietserkennung von Texten

Um herauszufinden, ob die Klassifikation von Texten durch Termini mit einer Klassifikation von Texten durch Texte übertroffen werden kann, wurde folgendes Experiment aufgesetzt:

In BISTRO wurden zunächst Fachglossare gebildet, also die Terme herausgefiltert, die ein Konzept der italienischen Rechtsordnung beschreiben, das nur einem einzigen Fachgebiet angehört. Da alle Terme mit Kontexten und Definitionen beschrieben werden, die besonders charakteristische Texte für den Term und sein Fachgebiet sind, ergeben die Beschreibungstexte zu Termen des selben Fachglossars ein kleines, virtuelles Fachkorpus. Der lexikalische Unterschied zwischen diesen Fachkorpora repräsentiert im Versuch den Unterschied zwischen den Rechtsgebieten selbst.

Aufgrund der lexikalischen Ähnlichkeit unbekannter Texte zu den Fachkorpora wird nun versucht, das Fachgebiet der Texte zu erkennen. Dazu betrachtet man jeweils 10 % der Textstücke als unbekannt und versucht anhand der verbleibenden 90 % ihr Fachgebiet zu erkennen. Dieser Test wird zehnmal mit anderen 10 % der Stichproben durchgeführt (zehnfache Kreuzvalidierung).[139]

---

[138] Eine weitere Voraussetzung ist, dass die Textsammlung repräsentativ für alle Texte des jeweiligen Fachgebiets ist.

[139] Die Kreuzvalidierung (*cross validation*) verallgemeinert einen Klassifikator, indem er ihn auch auf weitere Stichproben unbekannter Daten anwendet. Die Qualität der Klassifizierung wird aus dem Durchschnitt aller Versuche an den verschiedenen Stichproben ermittelt. So wird vermieden, dass die Klassifizierung an ein bestimmtes Trainingsbild angepasst wird und bei anderen Daten stets schlechter abschneidet.

Konkret erstellt man für jedes ‚unbekannte' Textstück (durchschnittliche Länge 388 Zeichen) und für jedes Fachgebiet jeweils eine Frequenztabelle. Das Fachgebiet mit der ähnlichsten Frequenztabelle wird dann als Fachgebiet des Textstücks vorgeschlagen. Ob es das richtige Fachgebiet ist, kann mit dem tatsächlich in der Datenbank vergebenen Fachgebiet überprüft werden. Aus allen Klassifizierungsversuchen lässt sich dann die gemittelte Korrektheit der Fachgebietserkennung berechnen. Die Ähnlichkeitsberechnung erfolgt über den Kosinus des Winkels zwischen den Merkmalsvektoren wie in Kapitel 1 Punkt 4 erläutert.

Es wurden in einer zehnfachen Kreuzvalidierung durchschnittlich 1127 Textsegmente mit einer gemittelten Länge von 388 Zeichen in 24 Fachgebiete automatisch eingeteilt. Das Trainingskorpus besteht insgesamt aus 10.061 Textsegmenten derselben Länge, ebenfalls in 24 Fachgebiete eingeteilt.

### 7.3. Vorversuche mit gemischten *n*-Grammen

In einem Vorversuch wurde nacheinander die Voraussagestärke der Frequenztabellen von Wörtern, 2-Grammen, 3-Grammen und 4-Grammen verglichen. Anschließend wurden entsprechend der Ergebnisse von Ogawa/Matsuda[140] und Ozawa et al.[141] auch *n*-Gramme unterschiedlicher Länge (*variable length n-grams*) kombiniert, und zwar 2- mit 3-Grammen, sowie 3- mit 4-Grammen.

Die Vorversuche zeigten, dass entgegen anders lautender Ergebnisse in der Fachliteratur[142] **gemischte *n*-Gramme** erheblich **schlechtere** Resultate brachten als die längere Sorte n-Gramme alleine. 3-Gramme alleine waren also deutlich besser als 2- und 3-Gramme gemeinsam, und 4-Gramme alleine waren deutlich besser als 3- und 4-Gramme gemeinsam, obwohl die Rechnerlaufzeit durch die Kombination natürlich stark erhöht wurde. Daher wurden gemischte *n*-Gramme nicht mit in den Hauptversuch aufgenommen.

### 7.4. Weitere Vorversuche

Im Vorexperiment zeigte sich, dass mit zunehmender Länge der *n*-Gramme die Wahrscheinlichkeit für Treffer als zweit- und drittähnlichstes Fachgebiet abnahm. Mit Ausdrücken des *Information Retrieval*[143] würde man sagen, dass bei starker Zunahme des *recall* gleichzeitig auch die Präzision gestiegen ist. Die zweite und dritte Wahl wurde daher nicht weiter untersucht.

Außerdem wurde festgestellt, dass die Erfolgsquote nicht sinkt, wenn alle einmal vorkommenden *n*- und *s*-Gramme (**Hapaxlegomena**) **nicht** in den Ähnlichkeitsvergleich mit einbezogen werden.

---

[140] Ogawa Y., Matsuda T. (1997), Overlapping statistical word indexing: a new indexing method for Japanese text, S. 226 – 234 in: Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval Philadelphia (USA) 1997, ACM Press New York 1997, http://portal.acm.org/ft_gateway.cfm?id=258576&type=pdf.

[141] Ozawa T., Yamamoto M., Umemura K., Church K. W. (1999), Japanese word segmentation using similarity measure for IR, S. 89-96 in: Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition (NTCIR Workshop 1), NACSIS (National Center for Science Information Systems) (Hrsg.) Tokyo 1999, erzeugten durch eine adaptive Methode *n*-Gramme verschiedener Länge, die bessere Ergebnisse brachten als *n*-Gramme gleicher Länge. Sie führten dies darauf zurück, dass 2-Gramme zu kurz seien, um fachsprachliche Texte korrekt abzubilden, http://research.nii.ac.jp/~ntcadm/workshop/OnlineProceedings/023-IR-Ozawa.pdf.

[142] Ogawa et al. a.a.O., Ozawa et al. a.a.O. Die von Ozawa et al. erzielten verbesserten Ergebnisse könnten z.T. darauf zurückzuführen sein, dass bereits die Segmentierung das Ähnlichkeitsmaß (TF/IDF) berücksichtigt, was voraussetzt, dass die Frequenz aller *n*-Gramme in allen Dokumenten berechnet werden muss. Übertragen auf die Fachgebietserkennung hieße das, dass die Klassifikationskraft aller n-Gramme für die Pseudofachkorpora berechnet werden müsste und dieses Wissen zur Segmentierung der zu erkennenden Textstücke verwendet würde. Bei einer zehnfachen Kreuzvalidierung müsste dies alles zehn Mal geschehen, sonst würden die Testtexte bereits in das Training eingehen.

[143] Wird meist mit Informationserschließung übersetzt, aber auch mit Abfrage, Informationsgewinnung, Informationsrückgewinnung, Wiederauffinden von Information.

Auch dieses Vorgehen spart viel Rechenzeit. Bei Wörtern wurde weiterhin mit den Hapaxlegomena gearbeitet, weil sie zur Erkennungsqualität beitragen.

In den Vorversuchen zeigte sich, dass Wort-*n*-Gramme durchaus nicht so schlecht abschneiden, wie das die computerlinguistische Fachliteratur durch ihre Fokussierung auf *n*-Gramme insinuiert. Da Zweiwortketten besser waren als Wörter, wurden im Experiment Ein- bis Vierwortketten untersucht, um den Zenit der Wörter im Untersuchungsbereich abzubilden. Die Vorversuche überraschten für *n*-Gramme und *s*-Gramme ebenfalls insoweit, als die in der Literatur so häufig besprochenen und verwendeten 2-, 3- und 4-Gramme noch nicht im optimalen Bereich für die Erkennung lagen. Das Experiment wurde also für 3- bis 8-Gramme durchgeführt, um auf jeden Fall die *n*-Gramme optimaler Länge im Untersuchungsbereich zu haben. Durch die hohe Anzahl an *n*-Grammen waren zur Durchführung der Versuche mehrere Wochen Rechenzeit nötig.

Bei den Vorversuchen wurden die Hypothesen entwickelt, dass die Länge der Versuchstexte zunächst einen entscheidenden Einflussfaktor darstellt, dass aber wie beim ähnlichen Problem der Sprachenidentifizierung ab einer gewissen Textmenge Sättigung eintritt.

Hypothese 1: Mit zunehmender Länge der Versuchstexte nimmt der Erkennungserfolg zu.

Hypothese 2: Mit zunehmender Länge der Versuchstexte nähert sich der Erkennungserfolg asymptotisch einem Maximum an.

Wenn sich die Hypothese 2 bestätigt, dann wäre mit der notwendigen Textgröße für eine Sättigung zugleich eine Aussage über die notwendige Größe des Fachkorpus gemacht.

Eine weitere Hypothese setzt die Länge der *n*-Gramme in Bezug mit der Länge des Versuchstextes. Bei besonders kurzen Texten könnten kurze *n*-Gramme besonders geeignet sein, Frequenztabellen zu erzeugen, die miteinander Ähnlichkeit aufweisen. Längere *n*-Gramme würden hier selten übereinstimmen und oft einen Kosinus von Null erzeugen. Entsprechend müssten längere *n*-Gramme bei längeren Texten besser abschneiden.

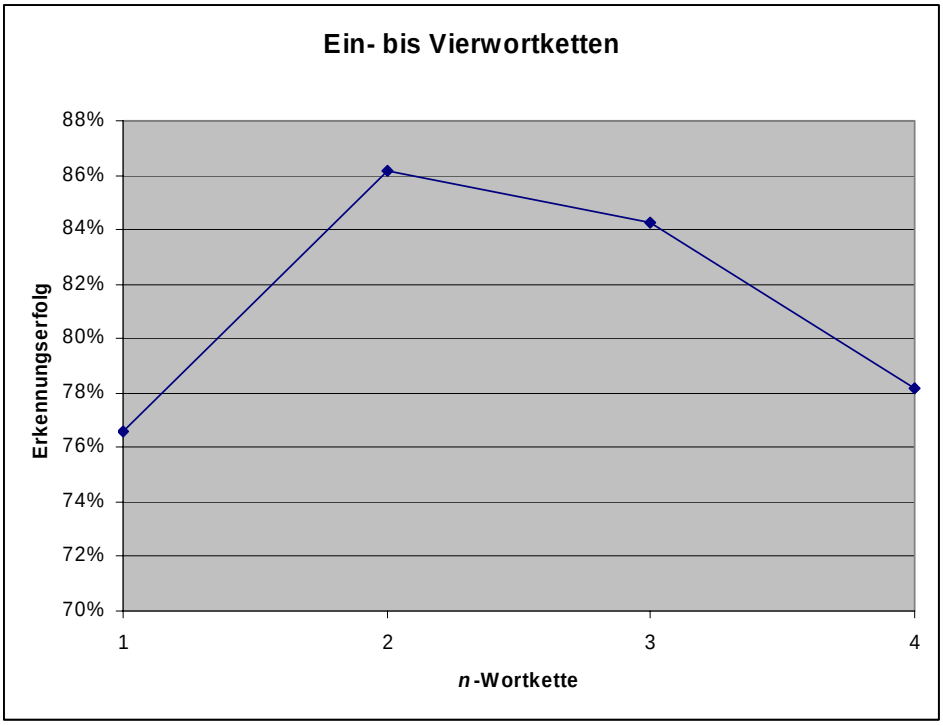
Hypothese 3: Mit zunehmender Länge der Versuchstexte verbessert sich der Erkennungserfolg von kürzeren *n*-Grammen im Vergleich zu längeren *n*-Grammen.

Zur Überprüfung der Hypothesen 1 bis 3 wurde die Länge der Texte, deren Rechtsgebiet zu erkennen war, variiert. Die Variation wurde präzise gesteuert, indem man die doppelte, vierfache, achtfache und zweiunddreißigfache Menge an Textstücken desselben Fachgebiets als eigene, längere Texte behandelte. Die Textstücke einfacher Länge wurden A genannt, die doppelt so langen (künstlichen) Texte B usw.

### 8. Fachgebietserkennung durch Textvergleich im Test

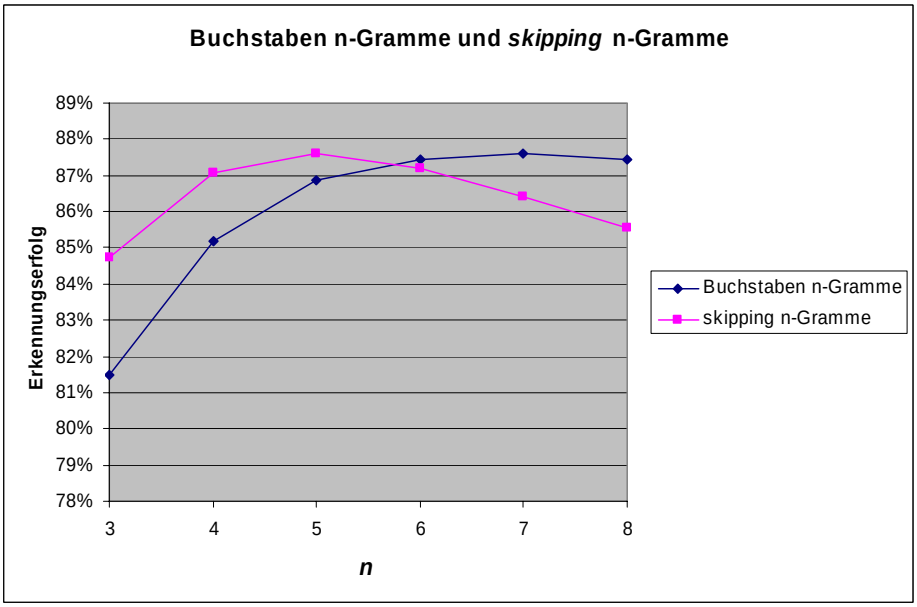
Für Wortgramme zeigt Grafik 12, dass Zweiwortketten das Fachgebiet der italienischen Texte am besten vorhersagen konnten.

**Grafik 12: Erkennungserfolg von 1-, 2-, 3- und 4-Wortketten**




*N*-Gramme und *s*-Gramme sind in Grafik 13 eingetragen. Die *s*-Gramme erreichten ihren Zenit bereits mit den 5-*skipping*-Grammen, die *n*-Gramme erst mit den 7-Grammen.

**Grafik 13: Erkennungserfolg von *n*-Grammen und *s*-Grammen**



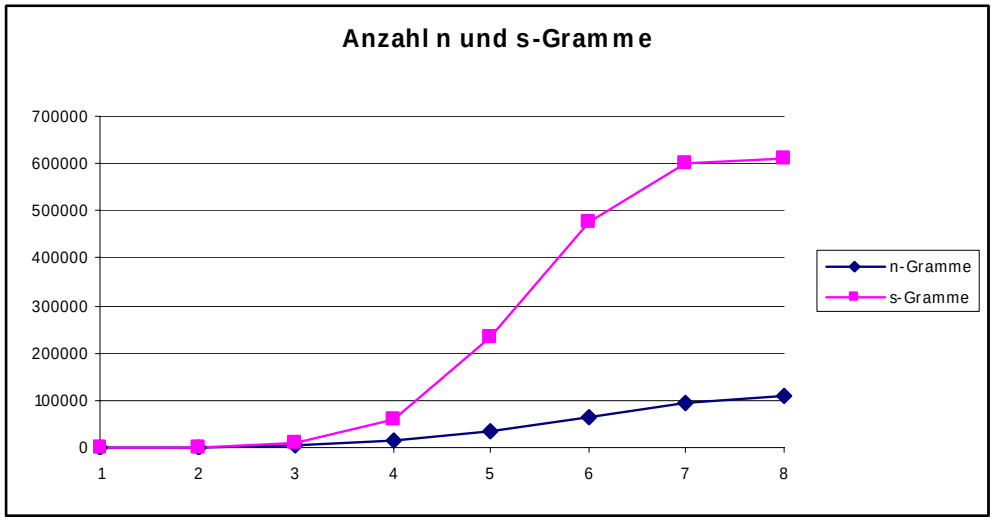

Bemerkenswert ist zunächst einmal, dass alle Ergebnisse im Bereich von nur 6 Prozentpunkten liegen. Die Wahl zwischen *n*- und *s*-Grammen und die Wahl der Länge haben offensichtlich nur für die Feinabstimmung eine Bedeutung.

Das bessere Abschneiden der kurzen ($n \leq 5$) *skipping* *n*-Gramme im Vergleich zu normalen *n*-Grammen gleicher Länge ist auf die größere Datenmenge zurückzuführen. Mit anderen Worten:

Die optimale Länge von *s*-Grammen ist kürzer als die von *n*-Grammen, weil bei gleicher Zeichenlänge mehr *s*-Gramme erzeugt werden als *n*-Gramme.

Der Zenit beider Kurven liegt aber nur wenige Prozentpunkte auseinander, denn die besten *n*-Gramme, nämlich 7-Gramme, übertrafen nur sehr knapp die besten *s*-Gramme, 5-*skipping*-Gramme. Um zu klären, warum *s*-Gramme ihren Vorsprung verlieren, wird die Anzahl der erzeugten *n*- und *s*-Gramme in Grafik 14 aufgezeichnet.

**Grafik 14: Anzahl erzeugte *n*-Gramme und *s*-Gramme**



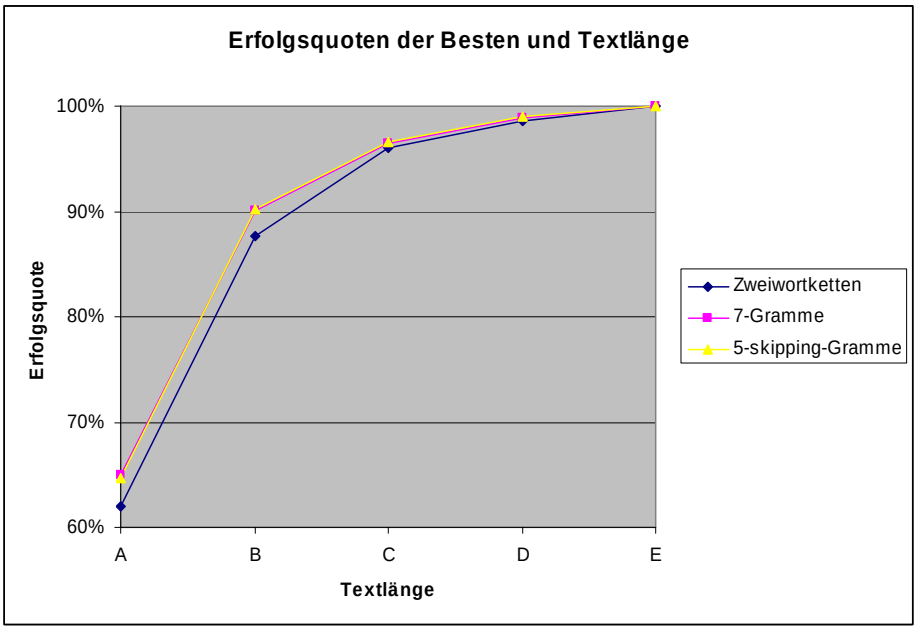

Interessant ist, wie sich die größere Kombinationsmöglichkeit (s.o.) auf die Anzahl der *s*-Gramme auswirkt: Es gibt erheblich mehr *s*-Gramme als *n*-Gramme derselben Länge. Bereits bei 3-Grammen werden 157 % mehr *s*-Gramme als *n*-Gramme erzeugt, 341 % mehr 4-Gramme und danach wird das Verhältnis noch ungünstiger für die *n*-Gramme. Der Erkennungserfolg hängt aber nicht nur von der Anzahl verschiedener Zeichenketten ab, denn 65.337 6-Gramme sind besser als 474.504 *skipping-6*-Gramme. Das spricht dafür, dass 6-Gramme eine größere Validität aufweisen. Daraus kann man schließen, dass das Optimum für die Erkennung zwischen den begrenzenden Faktoren ‚statistisch hinreichende Menge' für kleine *n* und ‚gute Validität' für große *n* liegt. Daraus folgt, dass sich der hohe Rechenaufwand, der sich für *s*-Gramme ab n>5 ergibt, nicht durch bessere Resultate gerechtfertigt ist, sondern dass ganz im Gegenteil die Ergebnisse schlechter werden.

Insgesamt deutet das Ergebnis darauf hin, dass *s*-Gramme entgegen anders lautender Hinweise in der Literatur keine bessere Vorhersage schaffen, sondern ihre Erfolgskurve entlang der x-Achse nach links verschieben.

Der Vergleich der absoluten Erfolgsquote zwischen Wörtern und *s*-Grammen zeigt, dass Wortketten mit gut 86 % den *n*- und *s*-Grammen mit etwa 87,5 % nur knapp unterlegen sind, obwohl die wenigen Wortketten erheblich weniger Rechenzeit benötigen.

Wenn man nun die Zeichenketten mit den besten Vorhersagewerten, also die Zweiwortketten, die 7-Gramme und die 5-*s*-Gramme über **verschiedene Textlängen** vergleicht, dann ergibt sich das Bild von Grafik 15:

**Grafik 15: Zweiwortketten, 7-Gramme und 5-*s*-Gramme**



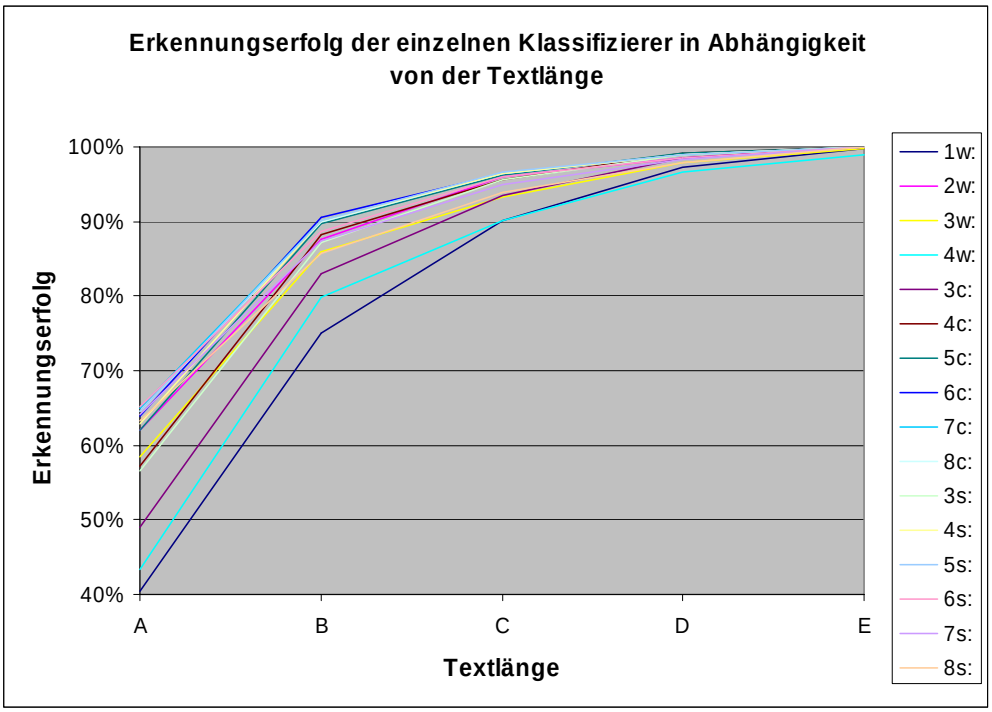

Die Erfolgsquote bei A in Grafik 15 stimmt nicht mit den Grafiken 12 und 13 überein, weil dort echte Kontexte in ihrer natürlichen variablem Länge verwendet wurden, während bei den Versuchen hier ohne Beachtung von Satz- oder Wortgrenzen jedes Textstück auf dieselbe Zeichenlänge gebracht wurde. Durch diese willkürliche Zerstückelung bzw. Anstückelung war der Erfolg zunächst niedriger. Diese Verfahren gewährleistete aber präzise Vervielfachung der Textlänge. Die Willkürlichkeit der Textgrenze fällt mit zunehmender Textlänge immer weniger ins Gewicht.

Bei allen Versuchen ergab sich ein klarer direkter Zusammenhang zwischen Textlänge und Vorhersageergebnis. Die *n*- und *s*-Gramme haben nur wenig mehr Erkennungserfolg als Wörter. Das gilt selbst bei kurzer Textlänge, wo Wörter sehr viel weniger Daten für den statistischen Vergleich bringen können. Die Ergebnisse für *n*- und *s*-Gramme sind annähernd identisch, was ein weiteres Indiz dafür ist, dass die Repräsentativität und Performanz der *n*- und *s*-Gramme auf dieselbe Ursache zurückgehen und sie sich daher ähnlicher sind als bisher angenommen.

Damit finden sich die **Hypothesen 1 und 2** eindrucksvoll **bestätigt**, weil der Erkennungserfolg stets von gut 60 % auf praktisch 100 % zunimmt. Nachdem die Fachtexterkennung durch Text praktisch perfekt funktioniert, ist dieser Ansatz der Erkennung durch Terme, deren Anfälligkeit oben qualitativ diskutiert wurde, klar überlegen. Nachdem Wörter als solche nicht schlechter abschneiden als *n*-Gramme (denn auch diese erreichen 100 %), könnte das an der Einschränkung auf Fachtermini liegen, durch die erst bei großen Textlängen hinreichend Daten für eine perfekte Erkennung vorliegen.

Die Hypothese 3 konnte nicht bestätigt werden, denn wie Grafik 16 deutlich zeigt, werden **alle Klassifikatoren mit Zunahme der Textlänge besser**. Damit ist die Regel ‚kurze *n*-Gramme für kurze Texte und lange für lange Texte' zumindest nicht allgemein gültig. In Anbetracht der exponentiell zunehmenden Rechenzeit bei längeren *n*-Grammen und größeren Textmengen sind daher stets die kurzen *n*-Gramme (oder gar die Wörter) zu empfehlen. Nur bei kurzen Texten wird man es sich erlauben können, um wenige Prozentpunkte der besseren Vorhersage zu kämpfen.

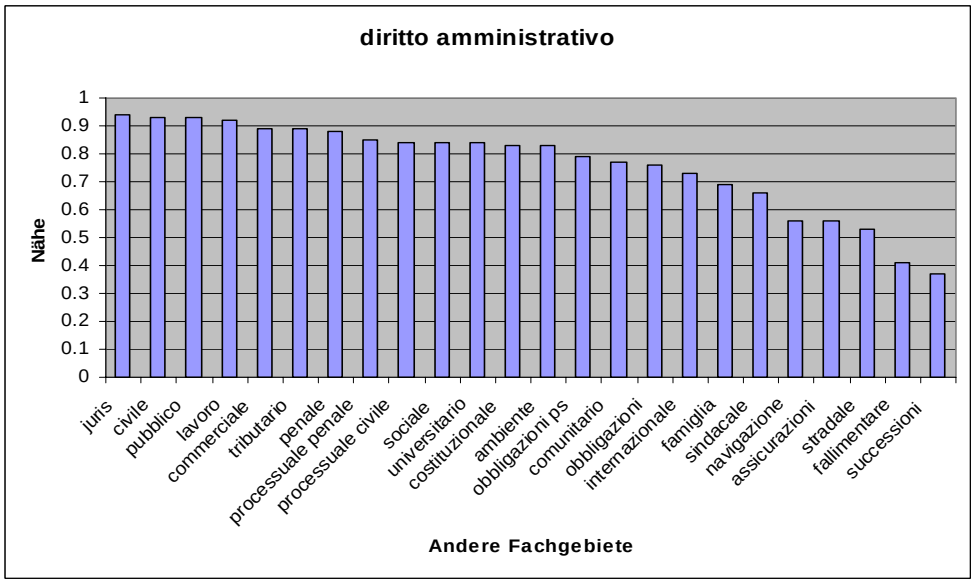
**Grafik 16: Alle Klassifikatoren**




## 9. Automatische Rekonstruktion der Fachgebietshierarchie

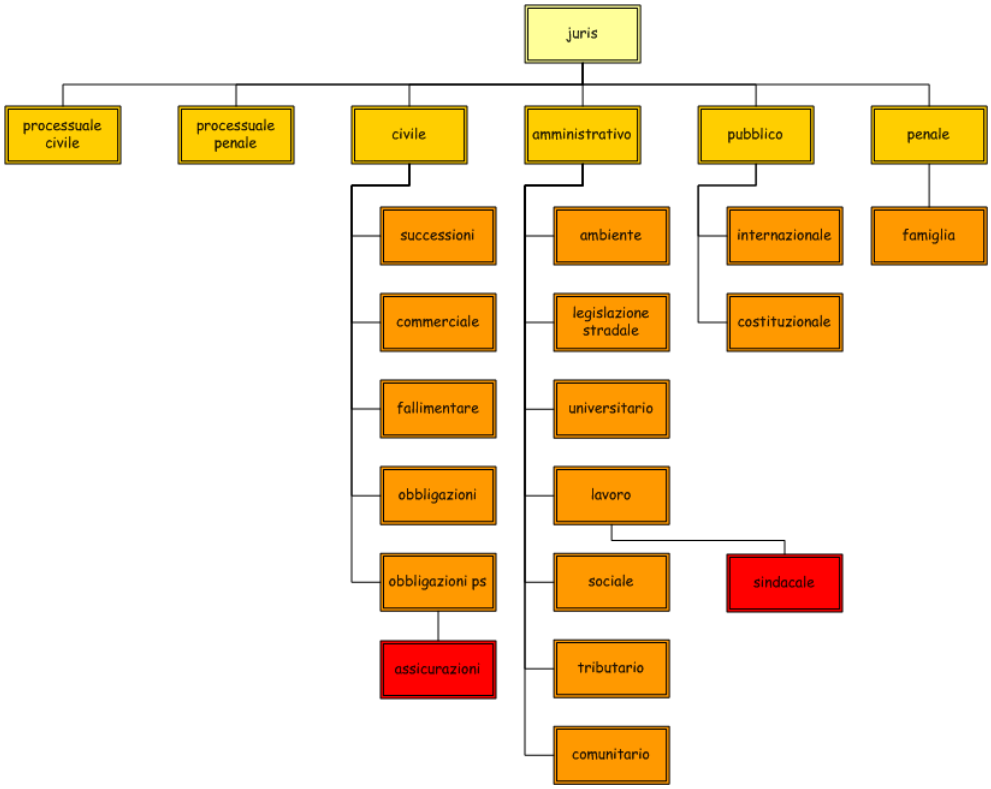
Mit der geschilderten Methode der Ähnlichkeitsberechnung lässt sich nicht nur das Fachgebiet von Texten erkennen, sondern auch die sprachlichen Nähebeziehungen der Fachgebiete zueinander berechnen. Hierzu werden alle Dokumente eines Fachgebiets als ein einziger Text betrachtet, die 24 Frequenztabellen der Fachgebiete erstellt und dann nach Kosinusähnlichkeit miteinander verglichen. Die Nähe des größten Fachgebiets Verwaltungsrecht zu allen anderen Fachgebieten wird in Grafik 17 gezeigt.

**Grafik 17: Ähnlichkeit des Verwaltungsrechts zu anderen Fachgebieten im italienischen Rechtssystem**



Es entsteht eine 24 x 24 Matrix mit 528 Feldern, welche die sprachliche Nähe jedes italienischen Fachgebiets zu jedem anderen angibt. Aus einer solchen Tabelle lässt sich durch ein geeignetes *Clustering* eine Ordnung der Fachgebiete erstellen und grafisch fassbar machen. Hierzu wird das Fachgebiet mit der höchsten gemittelten Ähnlichkeit zu allen anderen Fachgebieten als sprachlich zentrales Fachgebiet zur Wurzel des Hierarchiebaums. In die zweite Hierarchieebene kommen all die Fachgebiete, die diesem zentralen Fachgebiet näher stehen als irgendeinem anderen Fachgebiet. Für jedes dieser Fachgebiete werden dann wieder die Fachgebiete berechnet, die ihnen unter den verbliebenen Fachgebieten am nächsten stehen, usw. Hierdurch bekommen wir die Baumstruktur von Grafik 18, die die sprachliche Ähnlichkeit der Fachgebiete zueinander in einer Hierarchie widerspiegelt:

**Grafik 18: Sprachliche Verwandtschaft der Fachgebiete**




Übersetzung der Fachgebiete siehe Tabelle 5.

Zum einen ist beachtenswert, in wie weit die sprachliche Ähnlichkeit der Fachgebiete einer juristischen Klassifikation der Rechtsgebiete entspricht. Abweichungen zur juristischen Klassifikation können plausibel mit ihrer zumindest sprachlichen Nähe erklärt werden.[144]

Zum anderen ist bemerkenswert, dass zu diesem Clustering außer den Texten keinerlei sprachliche Informationen nötig waren. Für die Rekonstruktion des ‚fast richtigen' Fachgebietsbaums wurden weder Wörter noch Termini, noch Grammatik oder Flektionsanalyse benötigt. *N*-Gramme sind demnach trotz der instinktiven Vorbehalte, die einzelne Terminologen oder Linguisten ihnen gegenüber haben mögen, durchaus valide Indikatoren des semantischen Inhalts von Texten. Sie können Texte verschiedenster Art auf einfache Weise in eine hierarchische Beziehung setzen.

[144] Die beiden Fachgebiete im dunklen Kasten auf vierter Hierarchiestufe, „assicurazioni" und „sindacale", verwenden die am stärksten abweichende Sprache. Beim Versicherungsrecht könnten die wirtschaftlich-mathematischen Einflüsse stark sein, beim Gewerkschaftsrecht die politisch-korporativen. Beide werden jedenfalls richtig zugeordnet (Versicherungsrecht zum besonderen Schuldrecht und das kollektive Arbeitsrecht zum Arbeitsrecht).

Für solch ein Verfahren gibt es eine ganze Serie denkbarer Anwendungen. Für die Korpuslinguistik brächte es großen Nutzen beim Korpusaufbau (Automatisierung, Ausgewogenheit des Korpus). In der Terminografie könnte man durch diese Methode die wichtige Vorarbeit der Strukturierung des Arbeitsgebiets, also die Einteilung in Fachgebiete und ihre Zusammenhänge, zumindest entscheidend erleichtern. Eine weitere mögliche Anwendung dieses Clusterings in der Terminografie wäre die automatische oder semiautomatische Auswahl von sprachlich geeigneten, weil besonders prägnanten Kontexten aus Korpora oder Internetquellen, indem man zentrale Korpusstellen marginalen Korpusstellen vorzieht oder marginale Stellen absucht, um die Extension eines Terms zu ergründen.

## 10. Evaluierung der Fachgebietserkennung

Die Fachgebietseinteilung könnte mit anderen automatischen Verfahren verglichen werden, beispielsweise durch die Indizierung mit extrahierten Termen (Kapitel 3).

Die Richtigkeit der Ergebnisse der automatischen Fachgebietserkennung müssten allerdings mit direkter intellektueller Fachgebietszuweisung zu Texten verglichen werden. Dazu bräuchte man Texte, die eindeutig einem Fachgebiet angehören, was die in Punkt 3 geschilderten Probleme aufwirft. Im Experiment wurde diese Schwierigkeit umgangen, indem man für die Belegtexte der Datenbank ein Fachgebiet postulierte. Würden die Belegtexte explizit analysiert, dann wären sie losgelöst von ihrem Kontext im terminologischen Eintrag. Bei solch einer objektiven Interpretation durch Terminologen und Juristen würde bereits dadurch ein anderes Ergebnis erzielt werden. Die Abweichung durch fehlerhafte Zuordnung wäre nicht trennbar von der Abweichung durch die veränderte Aufgabenstellung.

Die intellektuelle Fachgebietszuweisung ist eine grundlegend andere Aufgabe: Beim Experiment wurde jeweils ein Text zu der ähnlichsten Textmenge zugeordnet, woraus sich implizit das Fachgebiet ergab. Intellektuell wäre ein Text dem passenden Fachgebietskonzept zuzuordnen. Dabei dürften Vergleichstexte keine Rolle spielen, denn sie könnten falsch zugeordnet oder untypisch sein. Die Aufgaben sind also etwa so unterschiedlich wie die Zuordnung eines Wortes zu Wortfamilien (automatische Erkennung) im Vergleich zur Definition der Bedeutung eines Wortes.

Außerdem wäre die Durchführung von intellektuellen Vergleichsversuchen sehr aufwändig, und da das Fachgebiet von der Textunterteilung abhängt (s.o.), müssten die Versuche wohl mit verschiedenen Textlängen gemacht werden.

Für die Praxis wichtiger wäre die Evaluierung, ob Terminografen mit Hilfe der Fachgebietserkennung nun tatsächlich besser oder schneller arbeiten. Das erforderte zwar einen komplizierten Versuchsaufbau und die Eliminierung von vielen Störfaktoren, wäre aber sicher ein besseres Argument für die Verbreitung von Fachgebietserkennern als die Erkenntnis, um wie viele Prozent die automatische Erkennung schlechter ist als die intellektuelle.

## 11. Zusammenfassung und Ausblick

Diese Untersuchung der computergestützten Fachgebietserkennung erbrachte folgende Ergebnisse:

- Texte eines Fachgebiets enthalten oft Textteile, die für sich allein genommen einem anderen Fachgebiet zugeordnet würden.
- Eine formale Unterteilung nach Kodifikationen kann nur wenige Hierarchiestufen abbilden und führt zu anderen Ergebnissen als die inhaltliche Beschreibung.
- Es gibt zwei Möglichkeiten, das Fachgebiet unbekannter Texte automatisch zu erkennen: 1. Absuchen auf Fachbegriffe, weil ein Text dem Fachgebiet angehört, dessen Fachbegriffe er benutzt. 2. Vergleich des Textes mit klassifizierten Texten, weil ähnliche Texte ähnliche Fachgebiete haben.

- Die Fachgebietserkennung durch Terme (1.) funktioniert besser, je mehr Termini in einem Text sind. Damit ist sie auch direkt von der Textlänge abhängig.
- Zur Fachgebietserkennung durch Fachtexte (2.) wird für jedes Fachgebiet ein kleines Korpus erstellt, aus dem eine Frequenztabelle aufgebaut wird, die als Vektor aufgefasst wird. Der Winkel zwischen Vektoren gibt die Fachgebietsähnlichkeit an.
- Die Fachgebietserkennung durch Fachtexte (2.) funktioniert wie die Sprachenidentifizierung annähernd perfekt. Die Erkennungsqualität korreliert direkt mit der Länge des zu erkennenden Textes.
- Der Erkennungserfolg kann durch die Länge der Vergleichsstücke (*n*-Gramme, skipping-*n*-Gramme oder Wörter) um einige Prozentpunkte optimiert werden.
- Mit der *n*-Gramm-Methode lässt sich nicht nur das Fachgebiet von Texten erkennen, sondern auch die sprachliche Nähe von Fachgebieten.
- Aus der Nähe aller Fachgebiete zueinander kann man eine Hierarchie der Fachgebiete clustern und grafisch darstellen. Diese Hierarchie kommt ohne sprachliche Informationen aus, reproduziert aber annähernd die intellektuelle Fachgebietseinteilung. Für solche Verfahren gibt es eine Reihe denkbarer Anwendungen.

Bei den vorgestellten Verfahren zur Fachgebietserkennung wird der **Nähe von Begriffen im Text** keinerlei Gewicht beigemessen, sofern sie nur im gleichen Textabschnitt zu finden sind. Die räumliche Nähe könnte aber zur Disambiguierung der Fachlichkeit beitragen, indem nur dasjenige der verschiedenen Fachgebiete für eine Benennung angenommen wird, das mit dem Fachgebiet der umgebenden Termini vereinbar ist. Steht die Benennung „Beihilfe“, die es sowohl im Strafrecht wie im Europarecht gibt, zwischen einer eindeutig verwaltungsrechtlichen und einer eindeutig europarechtlichen Benennung, dann kann darauf geschlossen werden, dass der Kontext von Beihilfe im europarechtlichen Sinne handelt. Dies kann die Fachgebietserkennung verbessern. Damit kann auch der richtige Wörterbucheintrag bei Homografen herausgefunden werden, ohne vom Benutzer weitere Information erfragen zu müssen und ohne ihn mit der Vieldeutigkeit des Begriffs konfrontieren zu müssen. Wer je eine Benennung im Wörterbuch gesucht hat und elf verschiedene Bedeutungen gefunden hat, wird das sehr zu schätzen wissen. Das Problem reduziert sich dann auf ‚nichtüberlappende Vektoren’, bei denen die Untersuchung der Nähebeziehung kein Ergebnis bringt.

Die Überprüfung der terminografischen Arbeit von Kollegen und insbesondere die Auswahl geeigneter Konzeptbeschreibungen durch Definitionen und Kontexte nehmen in der Praxis einen prominenten Anteil der terminografischen Arbeit ein. Durch die Fachgebietserkennung kann die Übereinstimmung von Texten mit dem Fachkorpus gemessen werden, so dass eine niedrige Einschlägigkeit oder Fachsprachlichkeit automatisch angezeigt werden kann. Das ist eine enorme **Arbeitshilfe für Terminografen**. Dieser Arbeitsschritt könnte von der Qualitätssicherung in die in die Produktion vorverlegt werden, denn dem Terminografen stünde bereits bei der Auswahl von geeigneten Beschreibungstexten ein Instrument zu ihrer objektiven Evaluierung zur Verfügung. Soweit abweichende Texte eine Aussage über ihre Qualität zulassen, kann die Fachgebietserkennung als Instrument zur ‚Selbstreinigung’ der Datenbank dienen.

Ein weiteres Anwendungsgebiet der automatischen Fachgebietserkennung ist natürlich der **Korpusaufbau** und die Korpuspräsentation. Das für jeden Benutzer zugängliche Korpus von BISTRO ist nach Sprachen getrennt, enthält aber keine Einteilung in Rechtsgebiete. Mit einer schnellen Fachgebietserkennung könnte gezielter im Korpus gesucht, könnten die Fundstellen nach Relevanz präsentiert und leichter interpretiert werden. Hinreichend große Texte könnten automatisch klassifiziert werden und das Korpus kann sich automatisch aufbauen, erneuern und erweitern.

Das bestehende Korpus könnte auf Fachsprachlichkeit geprüft werden, denn Gesetzestexte sind mitunter eher natur- oder sozialwissenschaftlich bis technisch statt rechtlich. Ungeeignete Texte

könnten ausgesondert werden und dafür gezielt nach objektiven Kriterien erweitert werden. Das Korpus gewänne damit die weitere Dimension der fachlichen Ausgewogenheit hinzu. Die **Gewichtung der Rechtsgebiete** könnte dann an das tatsächliche Vorkommen von Rechtstexten, an das tatsächliche Vorkommen von Fachgebieten in der Datenbank oder auch an die fachterminografischen Anforderungen angepasst werden. BISTRO enthält so viele Begriffe im Verwaltungsrecht, weil der terminografische Auftrag so lautete. Das Korpus spiegelt diese Anforderung derzeit aber nicht wider. Ziel könnte es daher sein, zu jeder Benennung der Datenbank auch einen fachspezifischen Kontext im Korpus zu haben.

Ein fachsprachlich gewichtetes Korpus böte dann auch einen besseren Ausgangspunkt für weitere Forschung etwa in der **automatischen Termextraktion**.

Ein weiteres Feld für Forschung wären die **Unterschiede innerhalb gleichsprachiger Rechtssysteme**. Die deutschsprachigen Benennungen der Datenbank BISTRO sind explizit dem italienischen, österreichischen, bundesdeutschen oder schweizerischen Rechtssystem zugeordnet. Unter Berücksichtigung der Unterschiede könnte die Fachgebietserkennung zu einer Rechtssystemerkennung verfeinert werden. Terminologen stehen oft vor dem Problem, dass die wenigen Belegstellen eines Rechtssystems in den vielen Belegstellen eines anderen Rechtssystems ‚untergehen', etwa die Südtiroler Benennungen in den gleich lautenden bundesdeutschen Benennungen. Bei der Suche nach Belegstellen im Internet bietet sich als pragmatische Lösung die Einschränkung der Suche auf ein Nationalitätskennzeichen im Domainnamen an (Suche bei GOOGLE nach Südtiroler Benennungen mit „site:.it"), was jedoch sehr unpräzise ist und alle Treffer von neutralen Domains (.org, .edu .gov usw.) ausschließt. Könnten die Treffertexte danach sortiert werden, wie viele Südtiroler Fachausdrücke sie enthalten, dann wäre dieses Problem gelöst.

In diesem Kapitel wurde gezeigt, wie die Terminografie von quantitativen computerlinguistischen Methoden profitieren kann, und zwar sowohl durch konkrete Arbeitserleichterung (Fachgebietsermittlung von Texten), bessere Ausnutzung der erzeugten Terminologie (automatische Erzeugung von Fachgebietsverwandtschaften) und neue Forschungsansätze (quantitative Fachsprachenforschung). Im nächsten Kapitel soll nun gezeigt werden, wie die Terminografie durch computertechnische Entwicklungen zu einer völlig neuen Konzeption gelangen kann.

# Literaturangaben zu Kapitel 2

Arntz R., Picht H., Mayer F. (2002), Einführung in die Terminologiearbeit, Studien zu Sprache und Technik Bd. 2, Georg Olms Verlag Hildesheim 2002.

Bergenholtz H., Tarp S. (Hrsg.) (1995), Manual of Specialised Lexicography – The preparation of specialized dictionaries, John Benjamins Amsterdam 1995.

Bortz J. (1993), Statistik für Sozialwissenschaftler, 4. Auflage Springer Berlin u.a. 1993.

Ciola B. (2001), Darstellung von Äquivalenzbeziehungen in der übersetzungsorientierten Terminologiearbeit im Recht, S. 742-752 in: Mayer F. (Hrsg.) (2001), Language for Special Purposes: Perspectives for the New Millennium, Bd. 2, Narr Tübingen 2001.

Gamper J. (1999), Construction of a Parallel Text Corpus Encoding Primary Data, in: Academia Nr: 18 (März - Juni 1999), EURAC, Bozen 1999, http://www.eurac.edu/Press/Academia/18/Art_13.asp :30.3.2004.

Grefenstette G. (1995), Comparing Two Language Identification Schemes, S. 263-268 in: Bolasco S., Lebart L., Salem A. (Hrsgg.) (1995), Proceedings of the 3. International Conference on Statistical Analysis of Textual Data (Journées d'Analyse de Données Textuelles JADT 95), CISU Rom 1995,
http://www.xrce.xerox.com/Publications/Attachments/1995-012/Gref---Comparing-two-language-identification-schemes.pdf : 8.4.2004.

Hahn W. v. (1983), Fachkommunikation: Entwicklung, linguistische Konzepte, betriebliche Beispiele, Sammlung Göschen 2223, de Gruyter Berlin, New York 1983.

Hoffmann L. (1975), Fachsprachen und Sprachstatistik. Beiträge zur angewandten Sprachwissenschaft, Akademieverlag Berlin 1975.

Langer S. (2002), Grenzen der Sprachenidentifizierung, S. 99-106 im Tagungsband KONVENS 2002, DFKI GmbH Saarbrücken, 2002.
http://konvens2002.dfki.de/cd/pdf/19V-langer.pdf : 27.1.2004.

Manning C. D., Schütze H. (1999), Foundations of Statistical Natural Language Processing, MIT Press Cambridge (USA), London 1999.

Mayer F. (2000), Terminographie im Recht: Probleme und Grenzen der Bozner Methode, S. 295-306 in Veronesi, D., Rechtslinguistik des Deutschen und Italienischen, Unipress Padova 2000.

Ogawa Y., Matsuda T. (1997), Overlapping statistical word indexing: a new indexing

method for Japanese text, S. 226 – 234 in: Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval, Philadelphia (USA) 1997, ACM Press New York 1997, http://portal.acm.org/ft_gateway.cfm?id=258576&type=pdf.

Ozawa T., Yamamoto M., Umemura K., Church K. W. (1999), Japanese word segmentation using similarity measure for IR, S. 89-96 in: Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition (NTCIR Workshop 1), NACSIS (National Center for Science Information Systems) (Hrsg.) Tokyo 1999, http://research.nii.ac.jp/~ntcadm/workshop/OnlineProceedings/023-IR-Ozawa.pdf.

Picht H., Schmitz K.-D. (Hrsgg.) (2001), Terminologie und Wissensordnung, TermNet Wien 2001.

Pirkola A., Keskustalo H., Leppänen E., Känsälä H., Järvelin K. (2002), Targeted s-gram matching: a novel n-gram matching technique for cross-and monolingual word form variants, in: Information Research 7 (2002): 2.

Schlohmann A. (1906), in: Illustrierte Technische Wörterbücher, zit. nach S. 366-367 in Picht a.a.O.

Streiter O., Voltmer L. (2003), Text Classification for Corpus-Based Legal Terminology, S. 253-260 in: Vrabie G., Turi J.G. (Hrsgg.), La théorie et la pratique des politiques linguistiques dans le monde, Tagungsband der 8. Internationalen Konferenz der Académie Internationale de Droit Linguistique 2002, Editura CUGETAREA Iaşi (Rumänien) 2003.

Internetquellen in Zitierreihenfolge:

http://www.cilta.unibo.it/Portale/RicercaLinguistica/bolc_eng.html : 21.6.2004.

http://www.cri.ensmp.fr/info_juridique/projet_info_juridique.html : 21.6.2004.

http://ontologie.w3sites.net/cgi-bin/codecouleur.pl : 21.6.2004.

http://www.xrce.xerox.com/competencies/content-analysis/tools/guesser : 21.6.2004.

http://search.admin.ch/cgi-bin/query?mss=de%2Fsimple&pg=q&what=web&enc=iso88592&fmt=.&q=Inverkehrbringen&submit=Suche: 30.1.2004.

http://www.scirus.com/ : 21.1.2004.

http://www.scirus.com/press/pdf/WhitePaper_Scirus.pdf: 30.1.2004.

http://InformationR.net/ir/7-2/paper126.html : 31.3.2004.

http://www.nicolaas.net/hapax/index.php? : 2.4.2004.

Kapitel 3

# III. Dynamische Termdarstellung

## Eine nutzer- und datenorientierte Darstellung von Rechtsbegriffen in einem elektronischen Rechtswörterbuch

Durch den Übergang von der physischen zur elektronischen Schrift erfährt die Rechtsterminologie und -terminografie große Veränderungen. Nunmehr muss nicht mehr eine Darstellungsweise allen Anforderungen gerecht werden, sondern die Darstellung kann dynamisch an die Umstände angepasst werden. In diesem Kapitel wird das Konzept einer dynamischen Termdarstellung vorgestellt und seine technische Umsetzung skizziert.

### 1. Zwei konkurrierende Paradigmen in der Terminografie

Terminografie ist die konzeptorientierte Fachsprachenbeschreibung für Fachleute durch Übersetzer und Fachgebietsexperten. In der Terminografie vollzieht sich ein Paradigmenwechsel. Die traditionelle Terminografie und die computergestützte Terminografie hatte sich von Ende der sechziger Jahre[145] bis Ende der Neunziger Jahre[146] einem absoluten oder statischen Modell verschrieben. Traditionell arbeitet die Terminografie mit festen Wissenseinheiten, die Einträge genannt werden und vor der computergestützten Terminologie eine eigene Karteikarte bildeten. Das terminologische Wissen sollte in jedem Eintrag dieselbe Struktur aufweisen. Daher definiert man vorab, welche Wissenselemente möglich sind und in welcher Beziehung sie zueinander stehen. Solch eine abstrakte Beschreibung der Datenkategorien heißt Eintragsmodell.

Die Einteilung terminologischen Wissens beruht auf der Annahme, dass die zu beschreibenden Wissenskonzepte bereits diese Einteilung aufweisen, die nur explizit auf Papier oder als elektronischer Eintrag reproduziert wird. Dieses traditionelle Paradigma findet weiterhin Anhänger in der Wissenschaft[147] und dominiert den Markt kommerzieller Anbieter von Terminologiemanagementanwendungen bzw. der Terminografiesoftware.[148]

Die traditionelle terminografische Einteilung vermittelt trotz erheblicher Komplexität innerhalb eines Eintrags (alle sprachlichen Bezeichnungen und deren Beschreibung gehören zu einem Konzept) eine einfache Orientierung und Kontrolle.

---

[145] Zu den ersten gehörten Brinkmann K.-H., Schulz J., Tanke E., (1969), Das Wörterbuch aus der Maschine, S. 9-15 in: Data Report 4/4 1969; nachgedruckt in: Graham J. D., Grewe K., Reisen U. (Hrsgg.) (1995), Terminologiearbeit. Theorie und Praxis, in: Festschrift für Eberhard Tanke zum 75. Geburtstag, Deutscher Terminologie-Tag Köln 1995.

[146] Mayer F. (1998), Eintragsmodelle für terminologische Datenbanken: ein Beitrag zur übersetzungsorientierten Terminographie, Forum für Fachsprachen-Forschung Bd. 44, Narr Tübingen 1998, spricht auf S. 222 von einem dynamischen Eintragsbegriff, der definiert wird als „Menge der Informationen zu einer Wissenseinheit, die ein Anwender für seine Zwecke aus dem Komplex der terminologischen Einheiten zusammenstellt".

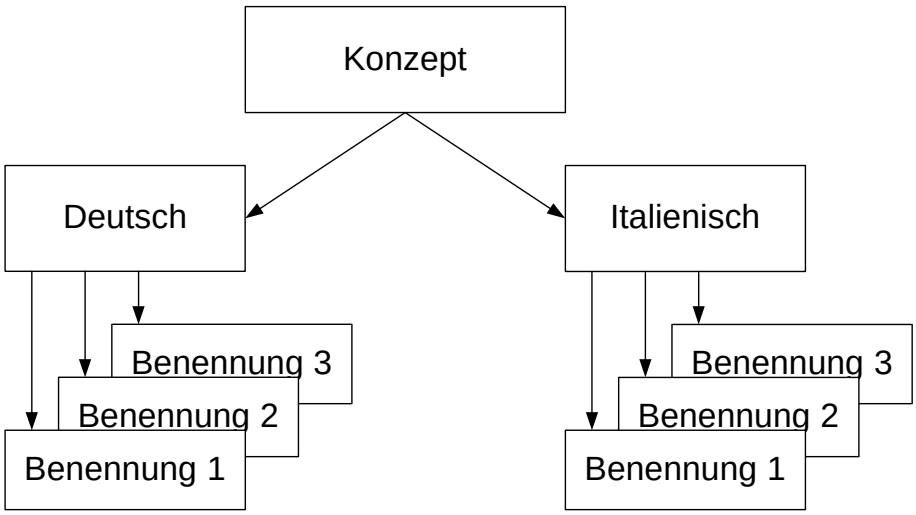
[147] Schmitz K.-D. (2002), Towards a uniform environment for representing terminologies within ISO, Präsentation auf der Terminology and Knowledge Engineering (TKE) Conference Nancy 2002, http://tke2002.loria.fr/Doc/workshops/ws2/ws2_kds.ppt : 18.3.2004.

[148] Z.B. Atril Deja-Vu, TRADOS Workbench, STAR Transit, SDL, SDLX. Vergl Liste unter http://www.uni-leipzig.de/~xlatio/software/soft-termiman.htm : 4.10.2004.

Das traditionelle Paradigma gerät aus drei Gründen unter Druck. Erstens gibt es viele verschiedene Eintragsmodelle,[149] was den Austausch von Terminologie erschwert und ein deutliches Anzeichen dafür ist, dass die Grundannahme einer objektiv erkennbaren Struktur des zugrunde liegenden Wissens ein bloßes Postulat ist. Zweitens ist jedes bisher erdachte Eintragsmodell an seine Grenzen gestoßen, weil es bestimmtes terminografisches Wissen nicht zufriedenstellend repräsentieren konnte.[150] Der dritte Grund ist, dass die EDV mittlerweile flexible Lösungen für die vielfältigen Anforderungen der Terminologienutzung bereitstellt, indem es die Datenverarbeitung (Erlangung, Speicherung, Modifizierung, Austausch) von ihrer Darstellung trennt. Diese drei Gründe werden nun eingehender erläutert.

Schmitz[151] schlägt zur Standardisierung aller Eintragsmodelle vor, dass jedes Konzept zunächst nach Sprachen zu unterteilen sei und anschließend nach Benennungen (Grafik 19).

**Grafik 19: Standardeintrag nach Schmitz**



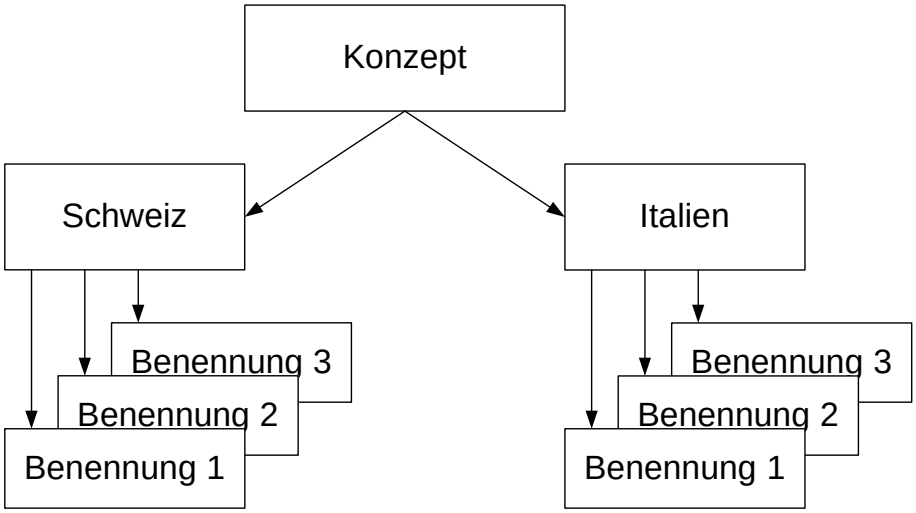

Diese intuitive Unterteilung erscheint zunächst einmal nachvollziehbar. Wenn man die dargestellte Struktur kommunikationstheoretisch analysiert, dann ist das Konzept das Thema, die Sprachen sind der Fokus und die Benennungen sind das Rhema. Die Besonderheit des objektiven oder statischen Paradigmas ist es nun aber, dass diese Struktur unveränderlich ist, also sowohl beim Speichern wie bei der Darstellung autoritativ ist. Kommunikationstheoretisch ausgedrückt entsteht eine monothematische, monofokussierte Kommunikation. Eine Darstellung der Daten, die den Fokus auf die Länder legt statt auf Sprachen und damit länderkontrastive Einblicke gewähren würde (Grafik 20), ist damit ausgeschlossen.

[149] Vergl. nur Mayer (1998) a.a.O.

[150] Vergl. z.B. Mayer F. (1996), The representation of inconsistent relationships in termbanks, S. 225-232 in: Galinski C., Schmitz K.-D. (Hrsgg.) (1996), Terminology and Knowledge Engineering Conference 1996 (TKE ’96), Indeks Frankfurt a. M. 1996.

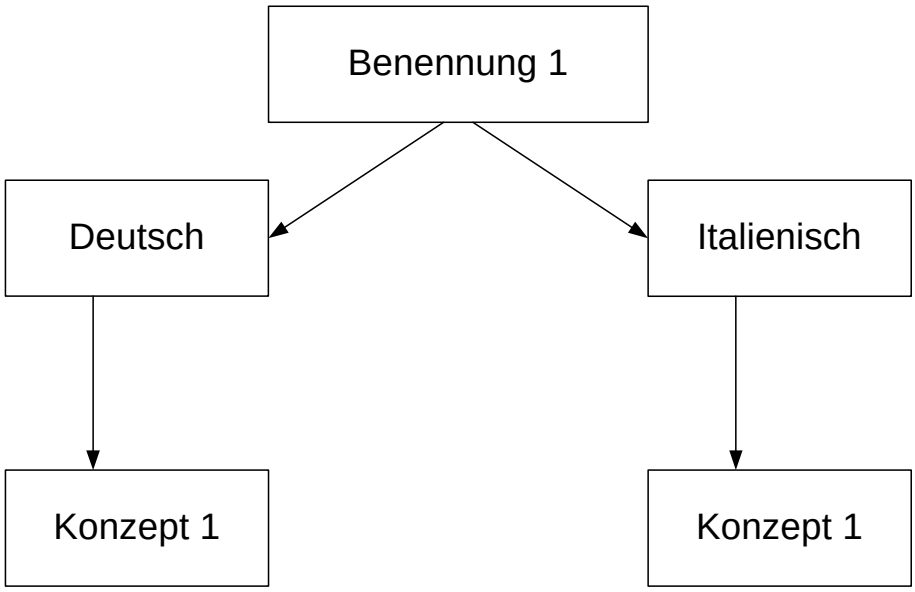
[151] Schmitz K.-D. (2002), Towards a uniform environment for representing terminologies within ISO, Präsentation auf der Terminology and Knowledge Engineering (TKE) Konferenz Nancy 2002, http://tke2002.loria.fr/Doc/workshops/ws2/ws2_kds.ppt : 18.3.2004.

**Grafik 20: Länderkontrastive Terminografie**



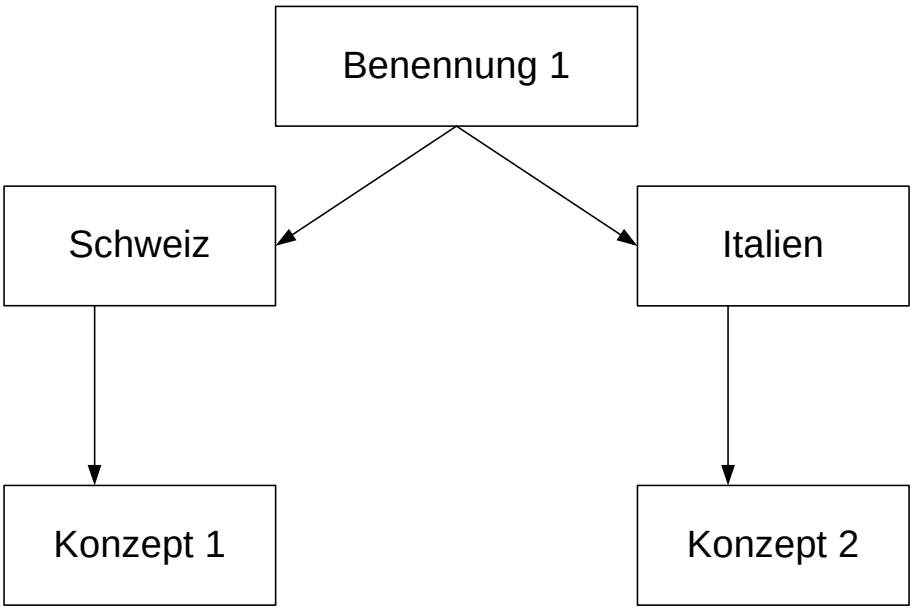

Zur Überprüfung von Einträgen könnte es auch interessant sein, die Daten nach einer Benennung als Thema zu sortieren und dadurch die mit homographischen Benennungen verbundenen Konzepte untersuchen zu können. Man könnte so falschen Freunden[152] auf die Spur kommen oder sprachkulturelle Unterschiede aufdecken (Grafiken 21 und 22).

**Grafik 21: Terminografie gegen falsche Freunde**




[152] (engl. *false friends*, frz. *faux amis*) Als falsche Freunde bezeichnet man vor allem morphologische und idiomatische Entsprechungen zwischen zwei Sprachen, wenn sich zwei Wörter oder Wendungen scheinbar entsprechen, aber unterschiedliche Referenzbereiche haben.“ Giese H., Kleppin K., Schlagwort ‚Falsche Freunde’, S. 204 in: Glück H (2000), Metzler Lexikon Sprache, 2. Auflage J. B. Metzler Verlag Stuttgart Weimar 2000. Siehe auch Beispiele unter http://en.wikipedia.org/wiki/List_of_false_cognates : 4.10.2004.

**Grafik 22: Sprachkontrastive Terminografie**




Ein Hinweis darauf, dass auch in der Praxis ein starkes Bedürfnis nach flexiblen **Thema–Fokus–Rhema**-Strukturen besteht, ergibt sich aus den so genannten Metatexten. Ein **Metatext** bietet Terminografen die Möglichkeit, in einem Freitextfeld diejenigen Informationen einzugeben, die in keine der vom Eintragsmodell vorgesehenen Datenkategorien passt. Eine Analyse von Hunderten dieser Metatexte[153] ergab, dass das Feld zur Disambiguierung verwendet wird, wenn im Standardmodell die Beziehungen der Benennungen untereinander nicht eindeutig sind. Als einfaches Beispiel sei hier das Konzept x in der einen Sprache mit A oder mit B benannt, in der anderen Sprache mit C und D. Das Standardeintragsmodell erlaubt nur perfekte Synonymie aller vier Benennungen. In Wirklichkeit kann es aber sein, dass A immer mit C zu übersetzen ist und B immer mit D, niemals aber A mit D oder B mit C, etwa aufgrund des Sprachniveaus, grammatikalischer Zwänge oder weil es so üblich ist.

Ein reales Beispiel aus der Rechtsterminografie ist der italienische Rechtsbegriff *contratto d'appalto*.[154] Diese Art von Verträgen entspricht im deutschen Recht u.a. dem Werkvertrag oder dem Dienstvertrag. Nach den Informationen des Eintrags zum Konzept *contratto d'appalto* müsste man davon ausgehen, dass für ein italienisches Konzept zwei Konzepte im deutschen Recht stehen.

[153] Dazu wurden alle Metatextfelder exportiert und intellektuell nach der Art der zusätzlichen Information geordnet. Nur selten wurde ein Feld wirklich für weiterführende Information genutzt. Meist bezog sich die Information auf die Beziehungen von Eintragsfeldern oder von Einträgen zueinander, die durch die Eintragsstruktur nicht eindeutig wiedergegeben werden konnten.

[154] Der Metatext lautet: „Im italienischen Recht steht der *contratto d'appalto* dem *contratto d'opera* sehr nahe; beide haben eine Leistung (Dienstleistung oder Werk) auf eigenes Risiko gegen Entgelt zum Inhalt, und der Auftragnehmer ist dem Auftraggeber nicht untergeordnet. Die Begriffe unterscheiden sich darin, dass der *prestatore d'opera* (einfache Werkunternehmer) vorwiegend eigene Arbeit leistet und daher Kleinunternehmer ist (Codice Civile=CC, Art. 2083), wohingegen der *appaltatore* (Unternehmer) vorwiegend im Rahmen eines Mittel- bzw. Großunternehmens auf fremde Arbeitsleistung zurückgreift (CC, Art. 2195, c. 1). Auch im deutschen und österreichischen Recht ist es nicht immer einfach, eine Abgrenzung zwischen Werk- und Dienstvertrag zu treffen. Im Gegensatz zum italienischen Recht erfolgt die Abgrenzung hier nach dem Inhalt des Schuldverhältnisses: Ist ein Erfolg geschuldet (z.B. die Herstellung eines körperlichen Werks), so ist es ein Werkvertrag; ist ein Tätigwerden geschuldet, handelt es sich um einen Dienstvertrag." Bullo F., Ciola B., Coluccia S., Maganzi Gioeni D'Angiò F., Mayer F., Treiber A., Voltmer L. (2003), Terminologisches Wörterbuch zum Vertragsrecht: italienisch/deutsch Dizionario terminologico del diritto dei contratti italiano – tedesco, C.H.Beck München, Athesia Bozen, Stämpfli Bern, Linde Wien 2003, S. 91-92.

Das ist aber falsch, denn der *contratto d'opera* entspricht im deutschen Recht ebenfalls manchmal einem Werkvertrag und manchmal einem Dienstvertrag. Damit besteht eben keine „eins zu zwei" Übersetzungsbeziehung, sondern eine Beziehung „zwei zu zwei anderen".

Im Grunde liegt in beiden Rechtssystemen ein noch abstrakteres, unbenanntes Konzept zugrunde (etwa: Leistung auf eigenes Risiko gegen Entgelt), das je nach Vorliegen oder Abwesenheit eines weiteren Tatbestandsmerkmals (in Italien: eigene oder fremde Arbeitsleistung; in Deutschland und Österreich: selbständige oder untergeordnete Arbeitsleistung) die eine oder die andere Benennung erhält. In diesem zusätzlichen Tatbestandsmerkmal unterscheiden sich die Rechtssysteme Italiens auf der einen Seite und Deutschlands und Österreichs auf der anderen Seite. Diese terminologische Situation kann im statischen Eintragsmodell nur durch die kontrastive Zusammenschau beider Einträge dargestellt werden; oder eben durch einen Metatext.

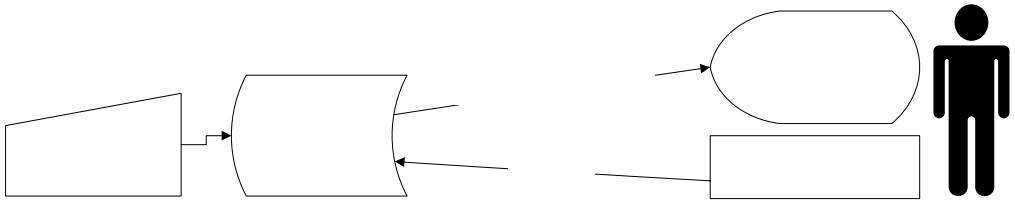
Auf den ersten Blick mag es übertrieben scheinen, wegen Spezialfällen ein bewährtes Eintragsmodell oder sogar die gesamte traditionelle Konzeption terminografischer Datenbanken in Frage zu stellen. Andererseits beweisen die Metatexte, dass es eben relativ häufig zu Spezialfällen kommt, die bei einheitlichen Eintragsmodellen unvermeidbar sind. Ein weiteres Indiz ist, dass moderne Terminologiesoftware zunehmend auch die Einbindung von Grafiken erlaubt, die nicht nur das einzelne Konzept,[155] sondern zunehmend auch Überblicksinformation zum Fach bieten.[156]

### 2. Technische Konzeption dynamischer Termdarstellung

Wie wird die dynamische Termdarstellung mit diesem Problem fertig? Datenverarbeitung und Datenpräsentation müssen getrennt werden und die Datenpräsentation muss die so gewonnene Freiheit zu einer kommunikativen Darstellung nutzen.[157]

Terminografie kann in drei Arbeitsschritte eingeteilt werden: 1) Datensammlung und Verwaltung 2) Datenspeicherung 3) Verbreitung der Ergebnisse (Grafik 23).

**Grafik 23: Drei Arbeitsschritte in der Terminografie**



Der Computer wurde in der Terminografie zunächst zur Unterstützung beim arbeitsaufwändigen Datensammeln, bei der Datenspeicherung und ihrer Verwaltung herangezogen. Beispiele sind die automatische Termextraktion[158], die Definition geeigneter Eintragsmodelle[159] oder die Datenver-

[155] Konzeptbeschreibend sind in einer Rechtsdatenbank etwa Grafiken von Straßenverkehrszeichen und -signalen, deren Bedeutung gesetzlich festgelegt ist.

[156] Wie Bullo F. et al. (2003) a.a.O. in Punkt XVII. Die terminologischen Diplomarbeiten des Instituts für Informationsmanagement der Fachhochschule Köln bieten nicht nur einen Begriffsplan, sondern sogar die Navigation über seine grafische Struktur: http://www.iim.fh-koeln.de/webterm/ : 4.10.2004.

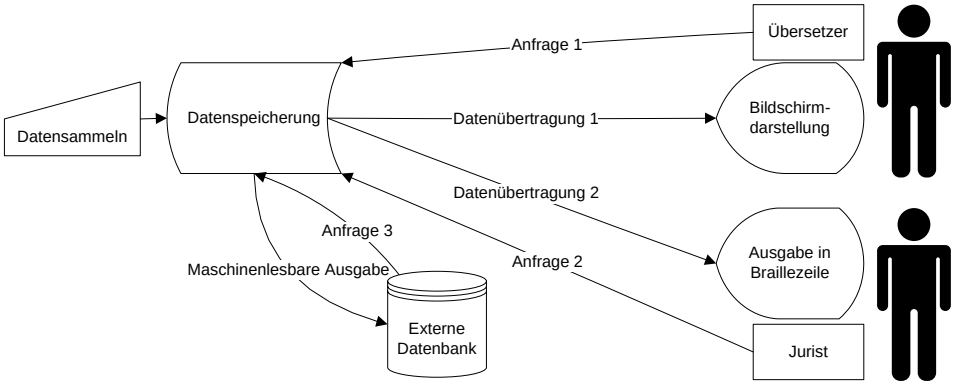
[157] Diesen Trennungsansatz mit anschließend generierter Datenansicht vertreten auch Sager J. C. (1990), A practical Course in Terminology Processing, John Benjamins Publishing Company Amsterdam 1990. Melby A. K., Wright S. A. (1999), Leveraging terminological data for use in conjunction with lexicographical resources, S. 544-569 in: Proceedings of the 5th International Congress on Terminology and Knowledge Engineering Conference Innsbruck (TKE '99), TermNet Innsbruck 1999, http://www.ttt.org/TKE-99.pdf : 25.3.2004.

[158] Vergl. Kapitel 4 oder Bourigault D., Jacquemin C., L'Homme M.-C., (Hrsgg.) (2001), Recent Advances in Computational Terminology, John Benjamins Publishing Company Amsterdam 2001.

waltung[160]. Die Datenpräsentation hat dagegen sehr viel weniger Forschungsinteresse angezogen. Die Gründe mögen darin zu suchen sein, dass die herkömmliche Weise der Darstellung von Terminologie (meist alphabetische Nachschlagewerke in Buchform) eine lange Tradition hat, die von Spezialisten für eine kleine und hoch spezialisierte Anwendergruppe erstellt und akzeptiert wird. Hinzu kam, dass mit der Entscheidung für ein Eintragsmodell praktisch auch die Entscheidung für eine bestimmte Darstellung getroffen ist.

Andersherum kann gesagt werden, dass die Verwirklichung verschiedener Sichten auf dieselben Daten auch ein Umdenken bei Datensammlung und -speicherung erfordern. Von einer anpassungsfähigen Darstellung profitieren insbesondere besondere Nutzergruppen (Hör- oder Sehbehinderte, Lese- oder Konzentrationsschwache, motorisch Behinderte, Computerneulinge) und besondere Darstellungsmedien (alter oder anderer Browser, langsame Datenverbindung, Sprachausgabe und Braillezeile statt Bildschirm, PDA[161] oder WAP[162]-Empfangsgerät, Touchscreen, direkte Datenübertragung in eine andere Datenbank). Die Terminologienutzer bekommen so schneller relevante und kohärente Information, so dass sie auch motivierter sein werden, mit dem System zu kommunizieren, das auf ihre Bedürfnisse, Interessen und Fähigkeiten eingeht. Die Auswahl der Daten und ihrer Übertragungsweise kann dabei auf den Modellen zur Nutzermodellierung und neuesten Internet- und Multimediatechnologien aufbauen.

**Grafik 24: Anpassungsfähige Darstellung**




Durch die Trennung der drei Arbeitsschritte können verschiedene Ansichten auf dieselben Daten dynamisch erzeugt werden. Bestimmte Aspekte der Daten können ausgeblendet werden, um die Komplexität zu reduzieren und so die Verständlichkeit zu erhöhen. In Grafik 24 wird für jede Anfrage eine eigene Darstellung erzeugt. Die Anfrage 3 kommt von einer externen Datenbank, die man am besten mit den rohen Daten bedient ist.

Wie bereits mehrfach betont, müssen Daten bereits anpassungsfähig gespeichert werden, um bei der Präsentation auf besondere Nutzergruppen, den Verlauf der Nutzerkommunikation und die ver-

---

[159] Mayer (1996) a.a.O.; Budin G. (2002), Der Zugang zu mehrsprachigen terminologischen Ressourcen – Probleme und Lösungsmöglichkeiten, in: In Mayer F., Schmitz K.-D., Zeumer J. (Hrsgg.), eTerminologie- Professionelle Terminologiearbeit im Zeitalter des Internet, Tagungsakte des Symposions "eTerminology", Deutscher Terminologie-Tag (DTT) Köln 2002. Melby A. K., Wright S. A. (1999), Leveraging terminological data for use in conjunction with lexicographical resources, S. 544-569 in: Proceedings of the 5th International Congress on Terminology and Knowledge Engineering Conference Innsbruck (TKE '99), TermNet Innsbruck 1999, http://www.ttt.org/TKE-99.pdf : 25.3.2004.

[160] Schmidt-Wigger, A. (1998), Building consistent terminologies, Poster in: Proceedings of the First Workshop on Computational Terminology (COMPUTERM'98) at the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics (COLING-ACL'98), ACL, Université de Montréal (Hrsgg.) Montreal 1998,
http://www.iai.uni-sb.de/docs/term2.pdf : 1.10.2004.

[161] PDA steht für *Personal Digital Assistant* und bezeichnet einen schnurlosen Handcomputer.

[162] WAP ist das *Wireless Application Protocol*, der Standard für kabellosen Internetzugang, z.B. mit Mobiltelefonen. Näheres zum *Wireless Application Protocol* (WAP) unter http://www.w3.org/TR/wbxml/ : 4.10.2004.

wendeten Medien eingehen zu können. Wie ist dies **konzeptionell** möglich (ohne Datenmodell? flexibles Datenmodell?) und ist dies **technisch** bereits machbar?

### 3. Datenspeicherung im terminologischen Datennetz

Dynamische Termpräsentation in Kommunikation mit dem Termnutzer baut auf diskursorientierten Modellen zur Generierung natürlicher Sprache auf, die auf den linguistischen Funktionalismus (Prager Schule) zurückgehen. Die Theorie wurde einerseits von Mel'čuk zur Meaning ⇔ Text-Theorie[163] weiterentwickelt, die der Textgenerierung bei maschineller Übersetzung zugrunde liegt,[164] andererseits von Halliday[165] zur systemisch-funktionalen Grammatik, auf der bedeutende Textgenerierungsprojekte aufbauen[166].

Insbesondere mit der Kohäsion und Kohärenz von Text beschäftigt sich die Rhetorische Strukturtheorie (*Rhetorical Structure Theory* - RST)[167], die damit vor allem die Generierung komplexer Texte beeinflusste.[168] All diesen Ansätzen ist gemeinsam, dass Datenansichten aus elementaren Datensätzen über deren Relationen generiert werden. Je kleiner die Informationseinheiten sind, umso komplexer aber auch flexibler ist das Zusammensetzen. Entsprechend komplex muss auch das Regelwerk zur Kombination der Wissenseinheiten sein. Die traditionelle Terminografie vermeidet diese Schwierigkeit, indem es auf Flexibilität verzichtet.

Das fundamentale Problem traditioneller Datenmodelle ist, dass sie jeden terminologischen Eintrag als eigenständige und von anderen Einträgen unabhängige Wissenseinheit konzipieren. Die Einträge selbst sind meist streng hierarchisch (‚linkseindeutig') aufgebaut, was sich in einer Baumstruktur mit gleichförmigen Unterknoten äußert. Die Alternative dazu ist ein nichthierarchisches Datenkontinuum, in dem jede Wissenseinheit mit jeder anderen verbunden sein kann. Auch in solch einem zyklischen Graphen sind die Wissenseinheiten (Knoten) und die prinzipiell zulässigen Beziehungen (Relationen) weiterhin präzise definiert. Es entsteht ein Netzwerk aus kleinsten terminologischen Wissenseinheiten, das in einer relationalen Datenbank gespeichert werden kann.

---

[163] Mel'čuk I. (2001), Communicative Organization in Natural Language: The Semantic - Communicative Structure of Sentences, Studies in Language Companion Series 57, John Benjamins Publishing Company Amsterdam 2001 mit weiteren Nachweisen. Vergl. allgemein die First International Conference on Meaning-Text Theory (MTT 2003), http://mtt2003.linguist.jussieu.fr/ : 29.3.2004 und die ETAP Systeme zur maschinellen Übersetzung.

[164] Apresjan J. D., Boguslavskij I. M., Iomdin L. L., Lazurskij A. V., Sannikov V. Z., Tsinman L. L. (1992), ETAP-2: The Linguistics of a Machine Translation System, S. 97-112 in: Meta, Bd. 37:1, http://www.erudit.org/revue/meta/1992/v37/n1/001895ar.pdf : 4.10.2004.

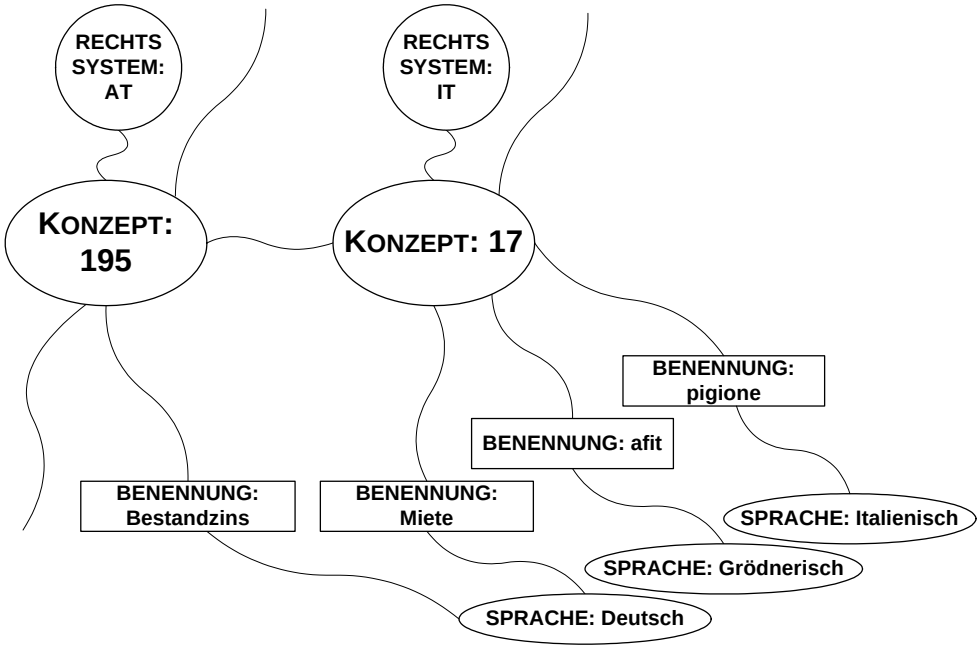
[165] Halliday M. A. K. (1994), An Introduction to Functional Grammar, 2. Aufl. Edward Arnold London 1994. Hallyday baut auf Firth und dieser auf Malinowski auf: Firth J. R., Palmer F. (Hrsgg.) (1968), Selected Papers of J.R. Firth 1952-59, Longman's Linguistic Library London 1968. Malinowski, B. (1923), The Problem of Meaning in Primitive Languages, Supplement I auf S. 296-336 in: Ogden C. K., Richards I. A. (Hrsgg.), The Meaning of Meaning, 8. Auflage Harcourt Brace & World New York 1946.

[166] Das Nigel/Penman System ist z.Z. wohl das größte und bedeutendste bisher implementierte Textgenerierungssystem für Englisch. Es wurde bereits vor 25 Jahren an der Universität Southern California am Information Sciences Institute (USC/ISI) entwickelt. Vergl. Matthiessen C., Bateman J. A. (1991), Text generation and Systemic-Functional Linguistics - Experiences from English and Japanese, Pinter Publishers London 1991.

[167] Mann W., Thompson, S. (1988), Rhetorical Structure Theory: toward a functional theory of text organization, S. 243-281 in: Text 8, 1988 (3).

[168] Hovy E. H. (1993), Automated Discourse Generation Using Discourse Structure Relations, S. 341-385 in : Artificial Intelligence (AI) 63 (1993): 1-2, http://citeseer.ist.psu.edu/hovy93automated.html : 29.3.2004.; De Carolis B., De Rosis F., Grasso F., Rossiello A., Berry D. C., Gillie T. (1996), Generating recipient-centered explanations about drug prescription, S. 123-145 in: Artificial Intelligence in Medicine, Vol. 8 (1996): 2.

**Grafik 25: Terminografie in einem Datennetz**




Die in Grafik 25 dargestellten Knoten sind RECHTSSYSTEM, KONZEPT, BENENNUNG und SPRACHE, durch die die Art des Inhalts definiert wird. Der konkrete Inhalt ist nach dem Doppelpunkt angegeben, wobei der Inhalt des Knotens KONZEPT nicht sprachlich ausgedrückt wird, sondern durch eine Identifikationsnummer. In gleicher Weise sind die verbindenden Relationen in ihrer Art definiert (in der Grafik nicht wiedergegeben).

Dieser beispielhafte Ausschnitt einer rechtsterminologischen Datenbank macht drei Vorteile deutlich.

1. Das Wissen kann präziser ausgedrückt und verwaltet werden, weil jeder Knoten und jede Beziehung unabhängig von einer Hierarchie, also unabhängig von anderen Knoten definiert ist. Im Beispiel hat jedes Rechtssystem seine eigenen Rechtskonzepte (KONZEPT 195 und KONZEPT 17), selbst wenn ihr Inhalt (momentan) genau dem Inhalt eines anderen Rechtskonzepts entsprechen sollte (was durch eine Äquivalenzbeziehung ausgedrückt wird).

In einem traditionellen Eintrag wird hingegen von einem Rechtskonzept ausgegangen, dem dann die Konzepte des anderen Rechtssystems untergeordnet sind. Ändert sich der Inhalt eines Konzepts, z.B. durch eine Gesetzesänderung, dann muss im traditionellen Eintrag die ganze Information zum geänderten Konzept von diesem Eintrag gelöscht werden. Im Datennetz genügt es bei einer Rechtsänderung, die Definition zu ändern und die Äquivalenzbeziehung zu korrigieren (zu entfernen oder evtl. einem anderen Konzept zuzuordnen).

Durch die unabhängige Definition von Knoten und Relationen vermeidet man implizite und mehrdeutige Beziehungen, die mit der Bezugnahme auf den Kontext eines Eintrags unweigerlich entstehen.[169] Hier kann wieder das Beispiel der Übersetzungsbeziehung *contratto d'appalto* zu Dienstvertrag und Werkvertrag angeführt werden, das auf jedem einzelnen Eintrag unvollständig

[169] Wenn in einem traditionellen Eintrag auf ein Feld BENENNUNG ein Feld QUELLE folgt, dann erkennt jeder Mensch an der unmittelbaren Aufeinanderfolge der Informationen, dass sich diese Quelle auf diese Benennung bezieht. Setzt man nun ein anderes Feld BENENNUNG zwischen die beiden Felder, dann ändert sich implizit die Beziehung des Feldes QUELLE. In einem Datennetz ist die Beziehung BENENNUNG-QUELLE explizit und kann sich niemals durch den Datenkontext (Hinzukommen oder Veränderung von an dieser Beziehung unbeteiligter Daten) ändern, sondern nur durch explizite Veränderung eben dieser Beziehung.

und irreführend ist, während es im terminologischen Netz zumindest schon einmal richtig gespeichert wird.[170] Es muss nun nur noch auf geeignete Weise dargestellt werden (dazu unten).

2. Es gibt keine redundanten Daten. Das KONZEPT 17 kann Miete benannt werden. Wenn nun auch das Konzept 195 Miete benannt werden kann, dann genügt die Erstellung einer Relation zu BENENNUNG: ‚Miete', während die Benennung in einem traditionellen Eintrag wiederholt werden muss. Die Vermeidung von Redundanz erleichtert dem Terminografen die Arbeit und hält die Datenflut im Zaum. In einer relationalen Datenbank heißt dieser Schritt Normalisierung.

3. Teilwissen kann gespeichert werden. Im traditionellen Eintragsmodell werden oft Minimalanforderungen gestellt wie z.B. dass jedes Konzept zumindest eine Benennung haben muss. Im relationalen Datenmodell kann ein Konzept auch nur mit seiner Definition in seinem Rechtssystem eingetragen werden. Dies genügt bereits für die Feststellung einer Äquivalenzbeziehung.

Wie unten zu zeigen sein wird, lassen sich diese drei Vorteile grundsätzlich auch realisieren, weil z.B. freie relationale Datenbanken hinreichend ausdrucksstark und technisch ausgereift sind, um zuverlässig terminologisches und lexikografisches Wissen in derselben Datenbank speichern zu können.[171]

## 4. Dynamische Termdarstellung aus dem Datennetz

### 4.1. Modell einer dynamischen Termdarstellung

Dynamische Termdarstellung impliziert, dass die Darstellung des terminografischen Wissens an den konkreten Nutzer angepasst wird, so wie Vokabular durch Grammatik an die konkrete Sprechsituation angepasst wird.[172] Traditionelle Termpräsentation baut von vornherein fertige Sätze, die Einträge, zusammen. Sie benötigt daher keine Grammatik, nutzt aber auch nicht den potentiellen Zugewinn an Ausdruckskraft z.B. durch Rekombination. Ein dynamisches Modell hingegen verknüpft inflektionsfähige Wissensbausteine durch variable Wortreihenfolge, Betonung und Thema−Rhema−Strukturen zu kompletten und sinnvollen Sätzen. Die in traditionellen Einträgen ein für allemal vorweggenommene Einteilung und damit Kontextualisierung des Wissens in Einträgen muss im dynamischen Modell erst zur Darstellung vorgenommen werden. Auf den ersten Blick scheint es schwierig, aus einem Wissenskontinuum einen nutzer- und kommunikationsorientierten sinnvollen Ausschnitt herauszuschneiden und einen kohärenten Diskurs zu bilden. An der Schwierigkeit dieser Aufgabe kann man erneut ermessen, wie arbiträr die Einteilung in Einträge ist. Andererseits muss die dynamische Termpräsentation als Mindestanforderung die Ausgabe eines traditionellen Eintrags ermöglichen. Eine weitere Mindestanforderung ist, dass die terminologischen Daten ganz ohne Darstellung übertragen werden können. Daraus erweitert sich das Dreistufenmodell einer anpassungsfähigen Terminografie wie in Grafik 26 dargestellt:

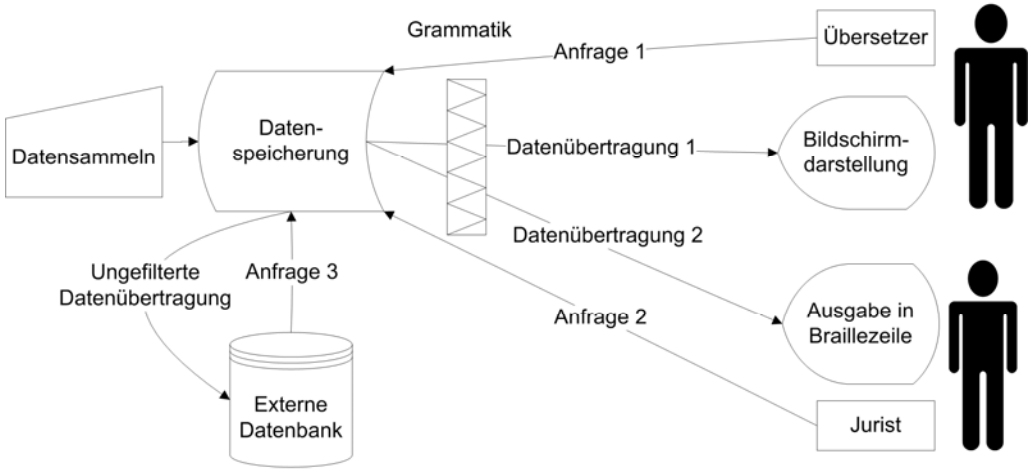
---

[170] „Richtig" speichern bedeutet explizit und inhaltlich speichern, wie das durch Speicherung in der *Extended Markup Language* (XML) der Fall ist. Die Felder in XML sind Layoutunabhängig beschrieben und verändern sich daher nicht mit dem Datenkontext.

[171] Holmes-Higgin P., Khurshid A. (1996), Is your Terminology in Safe Hands? Data Analysis, Data Modelling and Term Banks, Terminology and Knowledge Engineering, S. 215-224 in: Proceedings of the 4th International Congress on Terminology and Knowledge Engineering Wien (TKE '96), INDEKS-Verlag Frankfurt 1996.

[172] Der Vergleich mit einer Grammatik findet sich bereits in Streiter O., Voltmer L., A Model for Dynamic Term Presentation, S. 201-204 in: Tagungsband der TIA-2003 Konferenz Straßburg 2003, LIIA – ENSAIS, Université Marc Bloch (Hrsgg.) Strasbourg 2003, http://dev.eurac.edu:8080/autoren/publs/ModDynTerm.pdf : 4.10.2004.

**Grafik 26: Übertragung purer Daten alternativ zu Wissenspräsentation durch eine Grammatik**




Bereits mit der Erreichung dieses Modells wäre für die Terminografie viel gewonnen, denn die **Datenübertragung von Maschine zu Maschine** aus der terminologischen Datenbank ist frei von Layout, Formatierung oder Datenmodellspezifika. Konvertieren (Übertragen in ein anderes Datenformat), *Mapping* (Übertragen in andere Datenklassen) oder besondere Datenexportprozeduren (u.a. darf die Datenbank während des Exports nicht verändert werden und nur als Ganzes exportiert werden) entfallen.

Daneben passt eine Grammatik die Darstellung an **unterschiedliche Nutzerbedürfnisse** an, wie es in der Grafik 8 durch verschiedene technische Anzeigegeräte verdeutlicht wird. Spezielle Geräte der Nutzer (z.B. alternative Browser wie Opera, WAP, Braillezeile, Vorleseprogramme) erfordern die Trennung von Daten und Darstellungsbefehlen, wie es bei der Benutzung von *Style Sheets*[173] der Fall ist. Viele Nutzergruppen bekommen so erstmals direkten Zugang zu terminologischen Daten.[174]

Wenn terminologische Daten an maschinellen und an menschlichen Dialog angepasst werden können, dann können sie auch nahtlos mit der Darstellung in externen Termdatenbanken verknüpft werden. Der Nutzer bekommt dann für eine (Meta-)suche die Ergebnisse aus zwei Datenbanken. Das kann nur funktionieren, wenn die Daten für die gemeinsame Ergebnisliste in eine einheitliche Form gebracht werden können, wenn also der Inhalt von der Präsentation getrennt wird. Die Nutzer beider Termdatenbanken bekommen mehr Treffer und müssen sich nicht in eine fremde Datendarstellung eindenken.

Bereits die Anpassung der Daten an jegliches Medium und Ausgabegerät ist ein Fortschritt, die große Herausforderung ist aber die dynamische Generierung von terminologischen Diskursen wie in den Grafiken 1 bis 4. In Anlehnung an ein generatives Modell der Sprache werden hier die Metaphern einer Text-, Satz- und Wortgrammatik sowie einer Art ‚Artikulationsgrammatik' verwendet.[175]

---

[173] *Cascading Style Sheets*, siehe Fußn. 181.

[174] Das W3C gibt mit der *Web Accessibility Initiative* Empfehlungen, wie Daten gestaltet werden sollten, um auch diese spezifischen Darstellungsbedürfnisse befriedigen zu können: http://www.w3.org/WAI: 4.10.2004. Allgemein wird dieser Aspekt unter dem Stichwort *accessibility* behandelt.

[175] Zu einer ähnlichen Komplexitätsreduktion in der Textgenerierung und Textlinguistik vergl. Mel'čuk I. (2001), Communicative Organization in Natural Language: The Semantic - Communicative Structure of Sentences, Studies in Language Companion Series 57, John Benjamins Publishing Company Amsterdam 2001.

4.2. Beschreibung der vier Grammatikmodule

Die **Textgrammatik** regelt, wie der Dialog zwischen Mensch und Maschine über terminologisches Wissen generell abläuft. Traditionell ist der einzig mögliche Dialog eine Abfolge stets gleich zu formulierender Anfragen und die stets gleichförmige Ausgabe von Einträgen. In einem dynamischen Dialog kommen hingegen Textplanungstechniken zum Einsatz wie z.B. die Wiederholung des Substantivs, seine Wiederaufnahme in Oberbegriff oder Pronomen sowie der Verweis auf ein Wort oder einen Satz durch Adverb. Der Dialog Maschine-Mensch wird damit zu einer zielorientierten Kommunikation, in der die Intention der Sprecher (Erkennen der Anfragestrategie des Nutzers), der Textaufbau (Einleitung, Überleitung), rhetorische Formulierungsvarianten und allgemein die sprachliche Beziehung eine Rolle spielen. Dazu müssen Strukturen generiert werden, die über die einzelne Datenübertragung hinausgreifen.

Die **Satzgrammatik** regelt, welche Information auf den Ausgabebildschirm gelangen soll und stellt somit einen dynamischen terminografischen Eintrag her. Dieser ‚Satz' muss in den Dialog passen (dafür sorgt die Textgrammatik) und eine geeignete Antwort auf die Nutzeranfrage bieten. Die Satzgrammatik erstellt die hierarchischen und semantischen Strukturen zwischen den einzelnen Satzbestandteilen, also den Eintragsfeldern. Einem Übersetzer werden die Bezeichnungen nach Sprache sortiert, einem Juristen nach Rechtssystemen.

Die **Wortgrammatik** regelt, wie einzelne Bestandteile des Satzes verwendet werden, sie kontextualisiert also die Wissensknoten bzw. Eintragsfelder. Die Morphologie eines Wissensknotens wird der Sprechsituation angepasst, indem Muttersprachler nicht mit ausführlichen grammatikalischen Informationen und Juristen nicht mit langatmigen Abkürzungserklärungen belästigt werden, sondern mit einer Kurzversion. Die Flexion oder Auswahl einer Wortform kann auch das Layout miteinbeziehen, etwa wenn Rechtssysteme oder Sprachen mit einer Flagge symbolisiert werden.

Die **Artikulationsgrammatik** bestimmt das Layout der Ausgabedaten. Hier spielen vor allem Nutzerspezifika eine Rolle, etwa die Übertragungsweise, das benutzte Ausgabemedium und die individuellen Dysfunktionen auf Nutzerseite.[176]

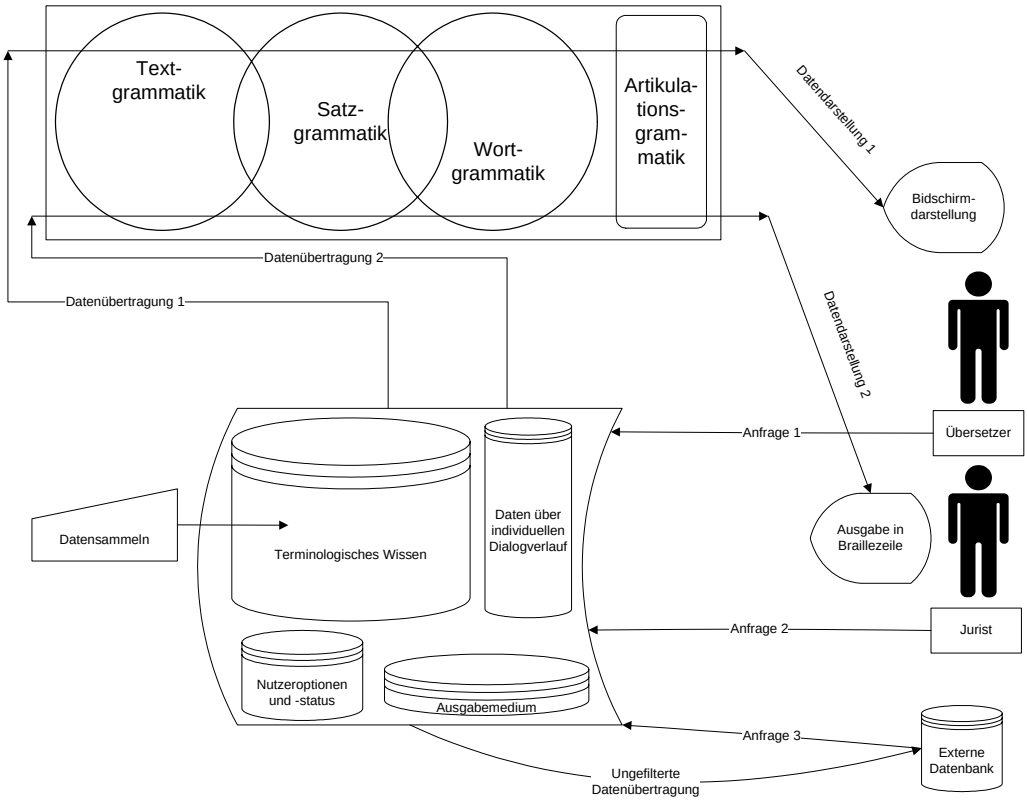
Die Information „italienische Benennung" kann z.B. durch die Position, den Zeichensatz, Laufweite, Schriftschnitt, -familie, -größe, -farbe, Hintergrundfarbe, animiert, akustisch und/oder grafisch, z.B. durch intuitiv erfassbare Symbole (Flaggen) dargestellt werden. Nicht alle Möglichkeiten werden von allen Benutzerschnittstellen in gleichem Maße und gleicher Weise unterstützt (Beispiel: keine Lautsprecher) und nicht jeder Benutzer ist gleich empfänglich für die verschiedenen Darstellungen (z.B. rot-grün Blinde), so dass eine Auswahl zwischen den Alternativen nötig ist.

Neben dieser Auswahl aus Alternativen ist aber auch eine kumulative Darstellung oft sinnvoll, z.B. durch eine Flagge und die Landesbezeichnung zusammen. Doppelte Informationen sind für Computer redundant bis gefährlich, für menschliche Benutzer haben sie aber einen pädagogisch-didaktischen Effekt, denn eine doppelte Information wird in aller Regel schneller, gründlicher und intuitiver aufgenommen. Daher sollten die vielfältigen Darstellungsmöglichkeiten genutzt und verschiedene Sinne angesprochen werden (Multimedia), um Nutzer möglichst stark zu aktivieren.

Die Wortgrammatik bekommt außerdem die Wünsche der Text-, Satz- und Wortgrammatik angetragen, die den konkreten Ausdruck durch Betonung bestimmter Satzglieder (die kontrastiven Felder hervorgehoben) beeinflussen möchten. Die vier Grammatiken interagieren also wie bei einer natürlichen Sprache (Grafik 27).

[176] Durch diese Anpassung werden Daten manchmal überhaupt erst zugänglich, siehe oben bei Fußn. 174.

**Grafik 27: Dynamische Termdarstellung mit vier Grammatikmodulen**




## 5. Techniken dynamischer Termdarstellung

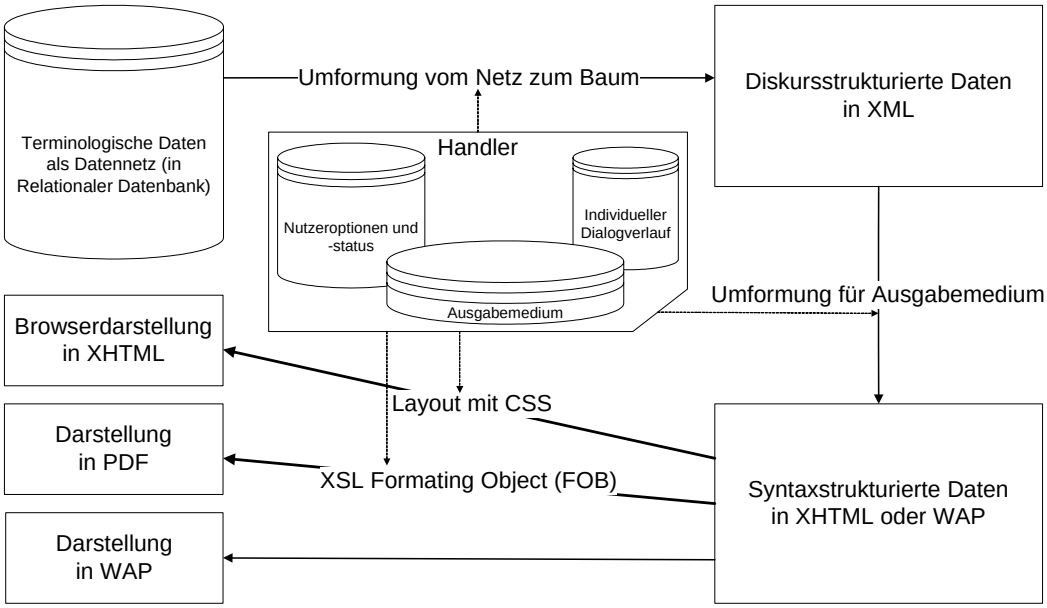
Dieser Abschnitt geht näher auf die technische Umsetzung der vier Grammatiken, auf Datenstrukturen und -umformungen und die Parameter der Darstellung ein.

Das Datennetz kann in einer relationalen Datenbank gespeichert werden. Die Knoten repräsentieren die Eintragsfelder wie KONZEPT, BENENNUNG, QUELLE, usw. Für solch elementare Datenkategorien gibt es Standards wie XCES und Dublin Core[177], ohne dass für die gesamte Datenbank ein eigener Standard eingehalten werden müsste. Diese Architektur garantiert größtmögliche Freiheit in der Datennutzung z.B. für Ein- und Mehrsprachige Wörterbücher, Fach- und Allgemeinwörterbücher, sowie für Wörterbücher mit oder ohne Definitionen, Kontexten, Grammatikangaben oder andern Datenkategorien.

Bei der Datenumformung von der relationalen Datenbank bis zur Darstellung sind drei Phasen zu unterscheiden: 1) Die Umformung terminologischen Wissens aus dem Datennetz zu einer ausschnitthaften, dialogstrukturierten **Baumstruktur**, 2) die ausgabemedienspezifische **Syntaxstrukturierung** des XML–Baums, 3) evtl. **Layouten**. Grafik 28 zeigt diese drei Phasen und den so genannten *Handler*, der die Umformung steuert.

[177] Dublin Core ist ein Metadatenset mit 15 Elementen, durch das elektronische Ressourcen besser beschrieben und aufgefunden werden sollen.

**Grafik 28: Drei Phasen dynamischer Termpräsentation**




In der ersten Unformungsphase bestimmen die Anfrage und weitere Parameter (dazu unten bei nächster Grafik), welche Knoten und Relationen aus dem Datennetz ausgeschnitten werden. Das geschieht durch SQL-Befehle[178] an die relationale Datenbank, deren zirkuläre Wissensstruktur dadurch in eine hierarchische XML-Wissensstruktur[179] umgeformt wird. Dabei erlauben textgrammatische Überlegungen die Auswahl aus den möglichen SQL-Befehlen zur Anfrage, z.B. die Thema-Rhema-Struktur, der Fokuskontrast und die Anbindung neuer Information an bereits früher ausgegebene Information.

Die Thema-Rhema-Struktur wählt das Ziel der Anfrage (Thema) und weitere anfragerelevante Informationen (Rhema) aus und trennt sie von allen übrigen irrelevanten Informationen. Durch den Fokus findet eine Kommunikationsabsicht Eingang in die Kommunikation. Die Maschine beantwortet die Anfrage mit einem Fokus, so dass beim Nutzer Interesse geweckt wird, den hervorgehobenen Punkt zu verstehen und im fortgesetzten Dialog mit der Maschine weiter zu vertiefen. Die Maschine führt den Nutzer von einem wichtigen Punkt zum anderen, statt sein Interesse durch die Abfolge stets gleichförmig aufgebauter und erschöpfender Einträge einzuschläfern. Die Maschine übernimmt damit einen weiteren Teil der intellektuellen Arbeit, nämlich das Trennen von kontextuell relevanter Information aus der enormen Menge damit zusammenhängender, aber momentan irrelevanter Information.

Für einen sinnvollen Diskurs ist auch die Kombination von Neuem mit Bekanntem wichtig. Nach einigen Anfragen braucht die Maschine keine Spekulationen mehr über den Nutzer anzustellen, sondern kann sich nach den bereits ausgegebenen Informationen richten, die hierzu gespeichert werden wie im Verlaufsordner eines Browsers (vergl. Daten über individuellen Dialogverlauf in Grafik 28).

Währen der zentrale Punkt als Thema einer Anfrage offensichtlich ist, bleibt für die rhematisch relevante Information meist viel Spielraum. Wenn keine besonderen Erkenntnisse vorliegen, dann

[178] SQL ist die *Structured Query Language*, eine ANSI- und ISO- Standardprogrammierungssprache zur Datenbankverwaltung und Anfrage.

[179] XML ist die *Extensible Markup Language*, ein vereinfachter Dialekt der *Standard Generalized Markup Language* (SGML), der Regeln für die Erstellung von spezielleren *markup languages* festlegt. Das bedeutet, dass ein eigenes Markup eingeführt werden kann, das trotzdem mit dem Metastandard XML kompatibel ist. http://www.w3.org/XML/ : 4.10.2004.

wird ein möglichst dichter terminologischer Raum ausgeleuchtet, indem die engsten Beziehungen des Themaknotens ausgegeben werden, wobei rekursive Beziehungen außer Acht bleiben.

In der zweiten Umformungsphase wird der dialogstrukturierte XML-Baum an das Ausgabemedium angepasst, indem aus dem XML das passende Datenformat erzeugt wird, z.B. (X)HTML, WAP oder PDF[180]. Da nicht in jedem Ausgabeformat dieselbe Menge Daten gut darstellbar ist (eine Handyanzeige ist viel kleiner), wird bereits die satz- und wortgrammatikalische Phase der ersten Umformung von der besten syntaktischen Struktur für das Ausgabeformat beeinflusst. Für die Datenansicht in Browsern steht mit *Cascading Style Sheets*[181] noch eine dritte Umformungsmöglichkeit für besondere Artikulationsweisen zur Verfügung, für das PDF-Format mit XSLT[182] *formatting objects* (FO)[183]. Für jede dieser drei Phasen gibt es also XML-Standardformate. Grafik 11 veranschaulicht die Auswirkung der Umformungen an einem Beispiel.

Die Technik und Standards zur Implementierung dynamischer Termdarstellung sind eine relationale Datenbank, XCES, Dublin Core, SQL, XML, XSLT, CSS und XSL-FO (vergl. Fußnoten 53, 177, 178, 179, 181, 182 und 183). Das Datennetz wurde deshalb in einer relationalen Datenbank statt direkt in XML implementiert, weil für relationale Datenbanken eine kostenlose, ausgereifte Arbeitsumgebung zur Verwaltung von Millionen von Daten bei kurzer Rechnerzeit zur Verfügung steht, während die Datenverwaltung in XML noch am Anfang ihrer Entwicklung steht.

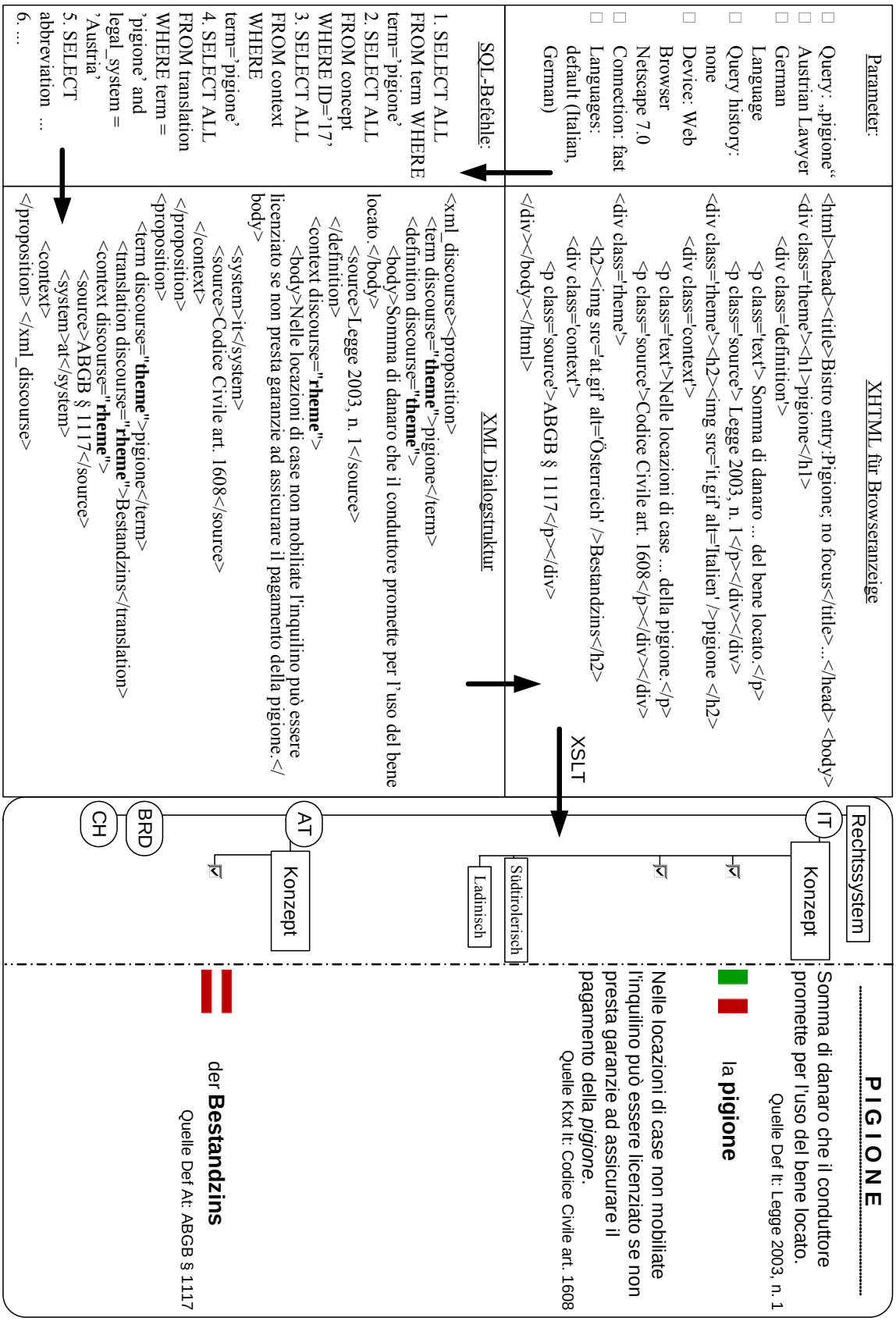
[180] Das von Adobe erfundene *Portable Document Format* (PDF) ist ein Datenformat, mit dem Dokumente unabhängig von Betriebssystem und Software identisch angezeigt wird.

[181] *Cascading Style Sheets* (CSS) sind der empfohlene Weg, um Dokumenten Layout hinzuzufügen. http://www.w3.org/Style/CSS/ : 4.10.2004. Es gibt auch *Aural Style Sheets*, mit denen Akustik in die Darstellung eingebunden werden kann: http://www.w3.org/TR/REC-CSS2/aural.html : 4.10.2004.

[182] XSLT, die *Extensible Stylesheet Language Transformation*, ist eine Programmiersprache zur Umkodierung von XML-Dokumenten, meist um das von Browsern anzeigbare XHTML zu erzeugen. Vergl. http://www.w3.org/XSLT : 4.10.2004.

[183] FO sind *formatting objects*. XSL FO ist ein XSL-Standard zur Umformung von XML durch ein XSL *Style Sheet* in eine andere Baumstruktur. Alles über *formatting objects* unter http://www.w3.org/TR/xsl/slice6.html : 4.10.2004.

**Grafik 29: Von Parametern mit SQL zu XML, XHTML und mit XSLT zu einer dynamischen Termdarstellung**



Grafik 29 zeigt die vier Grammatiken in der Umformung. Oben links sind als Parameter angegeben: a) Anfrage, b) Nutzerprofil mit Rechtssystem, Vorwissen und Muttersprache c) bisheriger Dialogverlauf, d) Informationen über das Ausgabegerät, e) Angaben über die Datenverbindung, f) Zielsprachen. (a) bestimmt das Thema, (b) entscheidet zwischen rechtsvergleichendem (Jurist) und sprachkontrastivem (Übersetzer) Rhema und wählt die Sprache der Benutzeroberfläche aus, (c) sichert die Kohärenz des Dialogs, (d) bestimmt die XML−Syntax, das Layout und z.T. die Größe einzelner Datenpakete, (e) bestimmt die Größe einzelner Datenpakete und z.T. das Layout, und (f) filtert die Zielsprachen aus den verfügbaren Sprachen heraus.

Die Parameter erzeugen SQL-Befehle (dem Pfeil folgend darunter) und holen aus der Datenbank: 1) das **Thema** ‚pigione', 2) die zutreffendsten Relationen nach den Parametern, 3) den KONTEXT zur Benennung, 4) rechtsvergleichende Information aus dem Österreichischen Rechtssystem. Die Ergebnisse der Datenbankabfrage werden zu einem XML−Baum geformt (dem Pfeil folgend im Fenster in der Mitte unten), der eine bestimmte Ansicht auf die terminologischen Daten bietet. Diese Diskursstruktur wird nun an das Ausgabegerät angepasst (Mitte oben), indem Begriffspläne, Grafiken (Flaggenzeichen o.ä.) und HTML-Text erzeugt und gestaltet werden. Die rechte Seite der Grafik zeigt einen möglichen Ausgabebildschirm.

Die Textgrammatik äußert sich entweder in einer Reihe dynamisch erstellter SQL-SELECT-Befehle wie im obigen Beispiel, oder in einem explizit vordefinierten SQL-VIEW[184], mit dem Eintragsmodelle präzise nachgeformt werden können. Der erste SELECT−Befehl holt den Themenknoten (hier ‚pigione') und fügt alle mit diesem Knoten verbundenen Relationen einem Kellerspeicher (*stack*) hinzu. Der zweite SELECT−Befehl verfolgt die Relation zum Fokusknoten (hier die Definition von ‚pigione') und holt alle Relationen dieses Knotens in den Kellerspeicher. Allerdings darf keine Relation zweimal verfolgt werden, sonst gerät das System in einen Zirkel. Aus den Parametern wird festgelegt, wie viele Knoten aus der Datenbank extrahiert werden sollen. Das Ergebnis dieser Befehle wird in eine XML-Struktur eingeschrieben, weil XML für die Umformung in jede Ausgabemodalität eine Sprache der XML-Familie bereithält.[185] Die Satzgrammatik wird als XSLT-Umformung implementiert, die je nach Ausgabegerät XHTML, WAP, XSL-FO oder SVG[186] erstellt. Die Wortgrammatik wird hauptsächlich bei der Auswahl aus alternativen XML-Attributen mit XSLT verwirklicht. Die 'Artikulationsgrammatik' kommt je nach Ausgabegerät entweder durch ein CSS oder die XSL-FO zum Ausdruck. In Browsern können *Style Sheets* auch abgewählt, überschrieben oder komplettiert werden, so dass der Nutzer in die Gestaltung miteinbezogen wird.

Nach dem Einsatz solch komplizierter Technik mag der Ausgabebildschirm auf den ersten Blick vielleicht etwas enttäuschend aussehen, so wie sich erst im Verlauf einer Konversation zeigt, wie gut das Gegenüber die Kommunikationstechniken beherrscht. Der Ausgabebildschirm ist jedenfalls eine individuelle Antwort auf die Anfrage des österreichischen Juristen nach der Bezeichnung ‚pigione'. Wenn der Nutzer nun detailliert nachfragt (das kann z.B. durch einen Klick auf die Definition geschehen), dann wird eine neue Anfrage mit einem Fokus ausgelöst, die mit einer Darstellung wie in Grafik 30 beantwortet wird. Auf der linken Seite wurde der Fokus auf das österreichische, auf der rechten Seite hingegen auf das italienische Rechtssystem gelegt.

Die beiden dynamischen Einträge für unterschiedliche Nutzer unterscheiden sich deutlich, obwohl der Aufbau beide Male dem traditionellem Eintragsaufbau der Terminografie folgt. Ebenso leicht lassen sich phantasiereichere Ausgabebildschirme erzeugen, wie der Begriffsplan und der sprachkontrastive Eintrag in Grafik 31 zeigen.

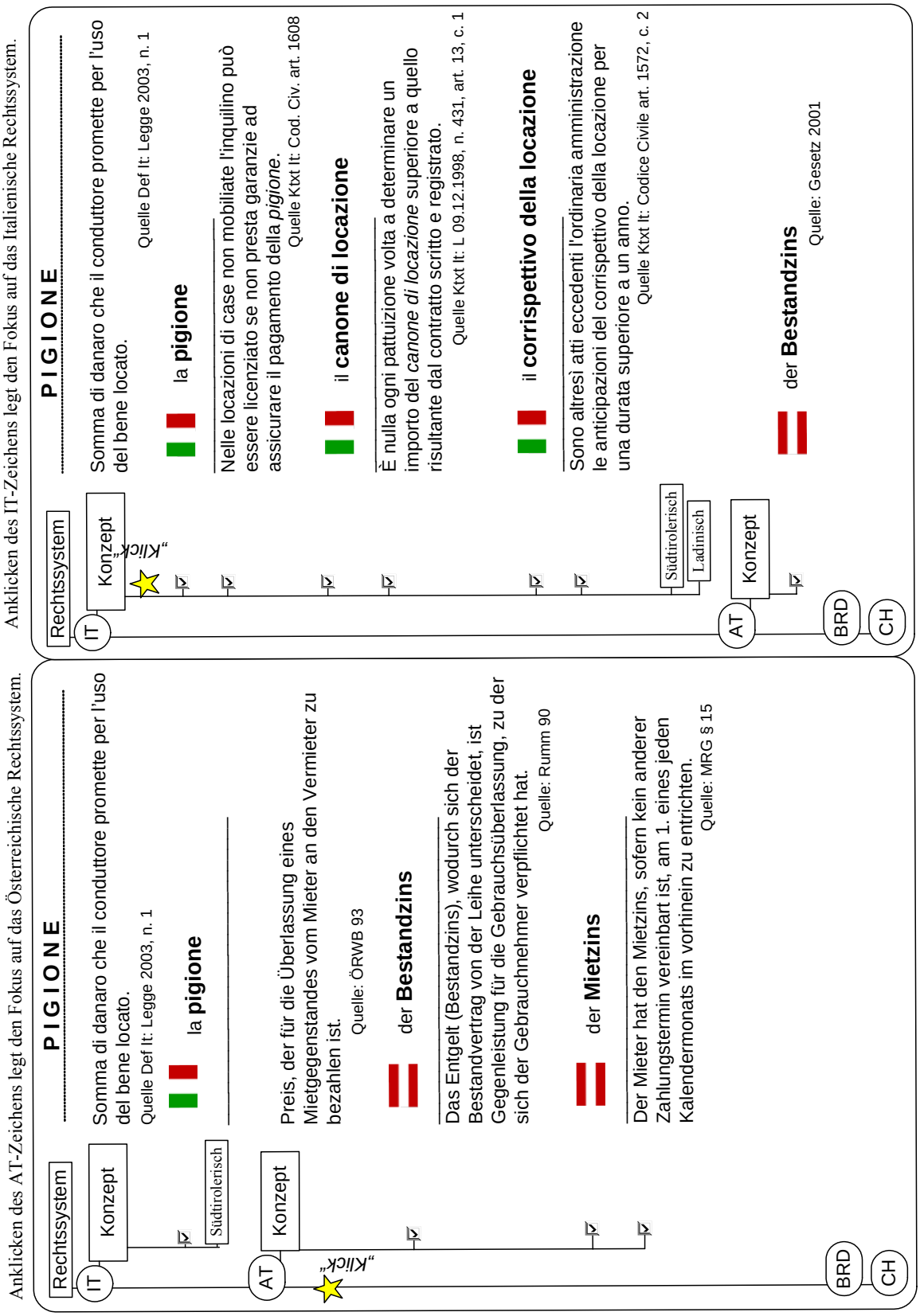
---

[184] So bei Ballew R., Duncan T., Blasingame M. (1999), Relational Data Structures for Implementing Thesauri, http://www.mip.berkeley.edu/mip/related/thesaurus/thesaurus.pdf : 4.10.2004.

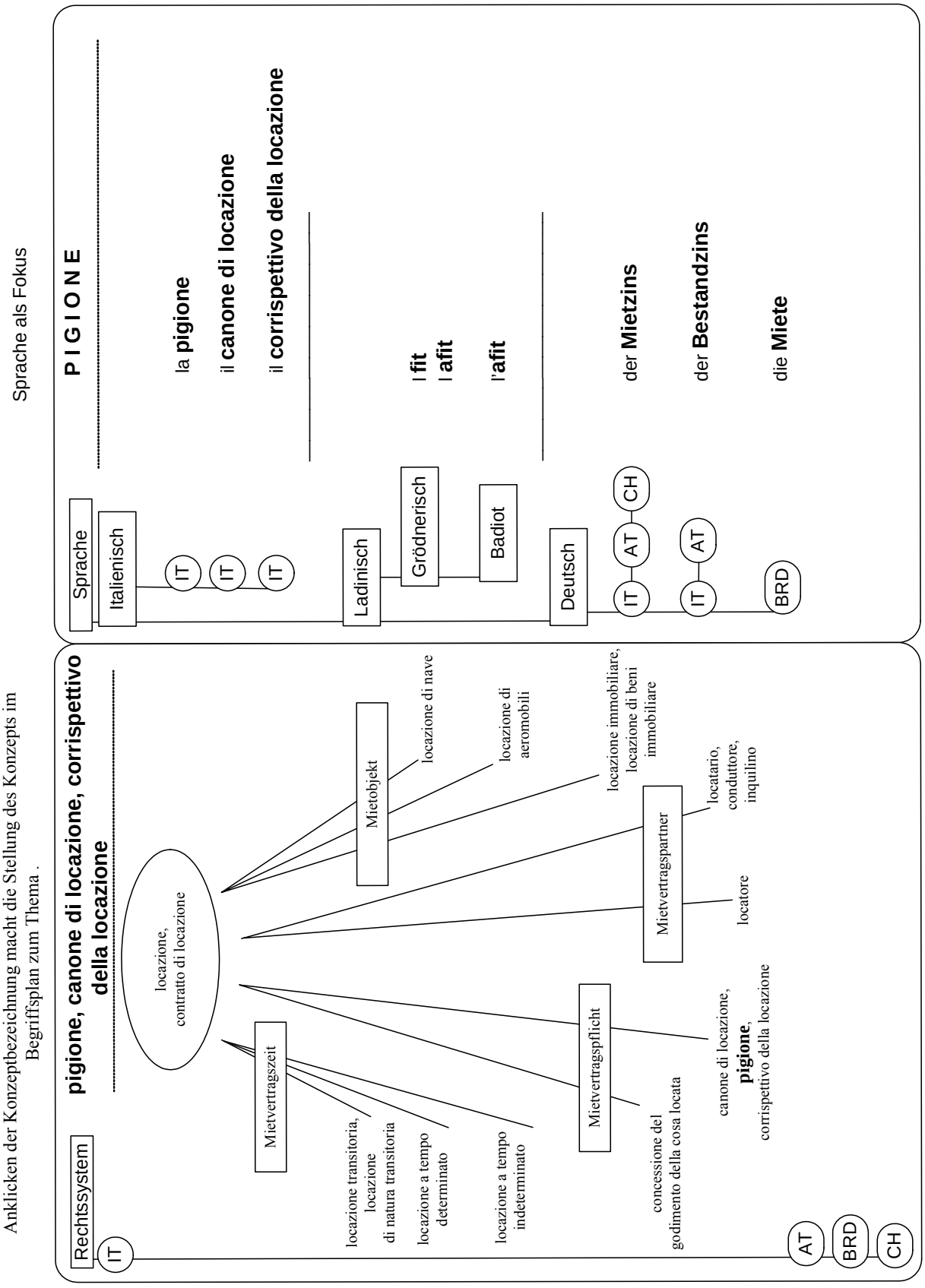
[185] Bourret R. (2003), XML and Databases, http://www.rpbourret.com/xml/XMLAndDatabases.htm : 4.10.2004.

[186] *Scalable Vector Graphics* (SVG) ist eine Vektorgrafiksprache im XML-Standard und wird vom W3C empfohlen.

**Grafik 30: Der vorige Eintrag mit Fokus auf die Rechtssysteme**



**Grafik 31: Begriffsplan und sprachkontrastiver Eintrag**



Die Ansicht auf der linken Seite zeigt eine automatisch erzeugte Vektorgrafik mit dem Begriffsplan für das italienische Rechtssystem. Da ‚pigione' keine Unterbegriffe in der Datenbank hat, werden der Überbegriff und dessen Unterbegriffe angezeigt. Die Unterscheidungskriterien, die auf den Strahlen als Kästchen gezeigt werden, sind an den deutschsprachigen Benutzer angepasst. Das Wort ‚pigione' bleibt weiterhin Thema und wird im Ausgabebildschirm hervorgehoben, Rhema sind nun die umgebenden Konzepte. Der Nutzer kann von diesem Überblicksbildschirm das Thema zu einem der angezeigten Konzepte oder anderen Rechtssysteme wechseln, z.B. durch Anklicken.

Die Ansicht der Datenbank auf der rechten Seite bleibt beim Thema ‚pigione' und fokussiert auf die Sprachen, so dass die Rechtssysteme den Sprachen untergeordnet werden. Aus dieser Ansicht erkennt der externe Nutzer unmittelbar, dass als deutschsprachige Übersetzung der Ausdruck ‚Mietzins' gebräuchlicher ist als ‚Bestandzins', und daher in einer Übersetzung oder für die Sprachnormierung vorzugswürdig ist. Der Terminograf erkennt aus der Ansicht, dass die Benennung ‚Mietzins' für das bundesdeutsche Rechtssystem noch nicht belegt ist. Auch in dieser Ansicht bieten sich durch Klicken auf die angezeigten Felder und Symbole eine ganze Reihe von Fortsetzungsmöglichkeiten für den Mensch-Maschine-Diskurs und für neue Anfragen und neue Ansichten der Datenbank.

## 6. Parameter dynamischer Termdarstellung

In Grafik 27 wurden neben terminologischen Daten und automatisch übermittelten Angaben über das Ausgabegerät auch zwei weitere Datenspeicher mit Daten über den individuellen Dialogverlauf des Nutzers, sowie mit dessen willkürlichen Einstellungen (Nutzeroptionen) und Berechtigungen (Status) gezeigt. Die nichtterminologischen Daten, die bei der dynamischen Termdarstellung verwendet werden, bilden für jeden Nutzer ein persönliches Nutzerprofil. Nutzerprofile werden nicht nur im elektronischen Handel, sondern auch in der Textgenerierung immer öfter verwendet.[187] Teilweise wird auch mit noch komplexeren Diskursplanungen experimentiert.[188] Der Diskurs soll mit diesen Methoden so gestaltet werden, dass das Kommunikationsziel erreicht (hier: rechtliches und sprachliches Wissen effektiv übermittelt) wird, wozu auch eine gewisse Bindung des Nutzers an das Instrument gehört.

Nutzerprofile für effektivere Kommunikation (und nicht zu Marketingzwecken) spiegeln immer die Optionen des Systems wider. Wenn es fachsprachliche und allgemeinsprachliche Definitionen in der Datenbank gibt, dann muss das in Grammatiken wirksame Nutzerprofil entscheiden, welcher Nutzer welche Definition erhalten soll. In gleicher Weise muss entschieden werden zwischen verschiedenen Datenkategorien (eine Definition für Muttersprachler oder viele Kontexte für Fremdsprachenlerner), zwischen der Anzahl Datenkategorien (nur Anfrage- und Zielbenennung, Zielbe-

---

[187] Beispielsweise bei der Präsentation von Kulturgütern für Webbrowser und Handgeräte, Ardissono L., Goy A., Petrone G., Segnan M., Torasso P. (2003), Intrigue: Personalized Recommendation Of Tourist Attractions For Desktop And Handset Devices, S. 687-714 in: Applied Artificial Intelligence: Special Issue on Artificial Intelligence for Cultural Heritage and Digital Libraries (2003) 17:8-9, http://www.di.unito.it/~liliana/EC/aai03.pdf : 29.3.2004.
Vergl. auch Peba-II, ein Online-Lexikon zur Beschreibung von Tieren, das aufgrund des Vorwissens der Nutzer und bisher abgerufener Information dynamische Vergleiche generiert, in denen zum besseren Verständnis Gemeinsamkeiten und Unterschiede zwischen bekannten und unbekannten Tieren hervorgehoben werden, http://www.dynamicmultimedia.com.au/peba/system.html : 4.10.2004. Diese Technik ist in das Dynamic Document Delivery System eingeflossen, http://www.cmis.csiro.au/iit/Projects/DDD/ : 4.10.2004.

[188] De Carolis B., Pizzutilo S., Palmisano I. (2003), D-ME: Personal Interaction in Smart Environments, S. 388 – 392 in: Lecture Notes in Computer Science, Vol. 2702, Springer-Verlag Heidelberg 2003, arbeiten mit autonomen Agenten, die alle am Diskurs beteiligten Interessen repräsentieren, miteinander verhandeln und so das optimale Resultat für die Datenanzeige bestimmen. Eine ähnliche Architektur mit Diskussion der verwandten Systeme ILEX und ARIANNA wurde bereits in De Carolis B. (1999), Generating Mixed-Initiative Hypertexts: A Reactive Approach, S. 71-78 in: Proceedings of the 1999 International Conference on Intelligent User Interfaces (IUI '99), IUI San Diego 1999, http://citeseer.ist.psu.edu/decarolis99generating.html : 4.10.2004 vorgestellt. Weiterführend insgesamt die International Conference on Intelligent User Interfaces, http://www.iuiconf.org/pastiui.html : 4.10.2004.

nennung mit allen Synonymen, Zielterm mit allen Relationen usw.) und zwischen verschiedenen Kommunikationsstrategien (Darstellung immer mit einem Konzept als Bezugspunkt oder auch Ausschnitte von Konzepten und Begriffspläne zur Übersicht).

Am einfachsten ist diese Entscheidung, wenn der Nutzer seine Präferenzen ausdrücklich angibt oder wenn sie wie die Art des Ausgabegeräts automatisch übermittelt werden[189]. Ansonsten muss einstweilen für den Nutzer gewählt werden. Den bisherigen Anfrageverlauf kann man nutzen, indem man typische Dialogmodelle festlegt und das tatsächlich Verhalten einem dieser Modelle zuordnet (z.B. durch ein *next-neighbour*-Verfahren, vergl. Kapitel 1 Punkt 4.9). Wenn das Nutzungsverhalten wesentlich abweicht, kann ein neues Nutzermodell angelegt werden, aus dem eine neue Dialogstruktur zur sinnvollen Darstellung terminografischen Wissens erkannt werden kann. Anstöße für die weitere Entwicklung des Systems kommen aus dem wachsenden Feld der Nutzermodellierung (*user-modeling*).[190] Eine in anderen elektronischen Wörterbüchern verwirklichte Idee[191] ist die Speicherung früherer Suchanfragen. Dadurch kann ein Benutzer seinen Weg durch die Datenbank explizit nachvollziehen, zu einem früheren Punkt zurückkehren, ausdrucken und evtl. bis zur nächsten Sitzung abspeichern.

### 7. Dynamische Darstellung in BISTRO

Terminologische Projekte sind meist auf lange Zeiträume angelegt und eine Umstellung der Datenspeicherung bringt in aller Regel enormen Aufwand bei der Bearbeitung der Altdaten mit sich. Auch das terminologische Projekt der EURAC läuft bereits seit über zehn Jahren. Nach Art. 99 des Sonderstatuts für die Region Trentino-Südtirol ist die deutsche Sprache mit der italienischen Amtssprache gleichgestellt, so dass eine rechtsverbindliche deutsche Version der Rechts- und Verwaltungsterminologie zu bestimmen ist. Dazu werden alle italienischen Rechtskonzepte von regionalem Belang aufgezeichnet und die entsprechenden Benennungen in italienischer und (soweit bereits vorhanden) deutscher Sprache aufgezeichnet. Dann werden rechtsvergleichend die Bezeichnungen entsprechender Konzepte der deutschsprachigen Rechtssysteme Österreich, Deutschland und Schweiz notiert. In einem dritten Schritt spricht eine Expertenkommission auf Vorschlag der EURAC eine Empfehlung an den Gesetzgeber aus, welche deutschsprachige Benennung für eine italienischsprachige Benennung zur Übersetzung festgelegt werden sollte. Dann beginnt der politische Prozess, an dessen Ende die gesetzliche Normierung der monodirektionalen Übersetzungsbeziehungen steht.

Nach Art. 102 des Sonderstatuts für die Region Trentino-Südtirol ist auch das Ladinische in gewissem Umfang zu fördern, so dass die terminologische Datenbank auch auf die Varianten des Ladinischen Grödnerisch, Gadertalisch und Fassanisch ausgedehnt wurde.

Die deskriptive and normative Arbeit wird in traditionellen Wörterbüchern[192] und in BISTRO veröffentlicht. Das terminologische Wissen wurde bisher mit der kommerziellen Terminologieverwaltungssoftware TRADOS aufgezeichnet und enthält neben den Benennungen: (a) Definitionen,

---

[189] Zur Einsicht in die zahlreichen Informationen, die bei einer Internetanfrage übermittelt werden siehe http://privacy.net/analyze/ : 4.10.2004.

[190] Ein verwandtes Projekt ist HyTex (http://www.hytex.info/ : 29.3.2004), in dem Korpora mit nutzerspezifischem Wissen automatisch annotiert werden, und zwar aus einem Wissensnetz mit Konzepten und Relationen heraus, angereichert durch textgrammatisches und linguistisches Markup. Beißwenger M., Storrer A., Runte M. (2003), Modellierung eines Terminologienetzes für das automatische Linking auf der Grundlage von WordNet, S. 95-104 in: Kunze C., Lemnitzer L., Wagner A. (Hrsgg.)(2003), Tagungsband des LDV-Forum 19 (2003): 1-2, Sonderheft mit Beiträgen des GermaNet Workshops zu Anwendungen des deutschen Wortnetzes in Theorie und Praxis.

[191] Le trésor de la langue française, http://atilf.atilf.fr : 4.4.2005.

[192] Zuletzt Bullo F. et al. (2003) a.a.O.

(b) Kontexte, (c) Quellenangaben, (d) die genormten Benennungspaare, und (e) mehrsprachige Parallelkorpora[193] zur Unterstützung der terminografischen Arbeit.

Um die über 15.000 traditionellen Einträge für die erläuterten Vorteile dynamischer Termpräsentation und -verwaltung umzubauen, mussten die Einträge in ein Datennetz umgewandelt und in einer relationalen Datenbank gespeichert werden. Hierzu wurden Felder und Relationen der Einträge analysiert und auf 25 verschiedene elementare Inhaltsknoten zurückgeführt[194] und in verknüpfte Tabellen überführt. Diese Umwandlung kann automatisch erfolgen und sofort durch Datenansichten (*data-views*) überprüft werden. In BISTRO können so ganze mehrsprachige Korpora samt Alinierung und Metadaten als Datenansicht dynamisch erzeugt werden.[195]

BISTRO unterstützt die intuitive Erkennbarkeit der Kommunikationsstruktur durch systematischen Aufbau der Datenansichten. So wird das Thema stets zuerst angeführt und vor dem Kontext des Rhemas (farblich) markiert. Es folgt wiederum (farblich) markiert der Fokus und schließlich einige explizit angegebene Vorschläge zum Themenwechsel und zur weiterführenden Abfrage von Daten, also etwa die Anzeige der geordneten Fundstellen in den Parallelkorpora (*keyword-in-context* KWIC) oder die gezielte Suche in ausgewählten Datenbanken des WWW.

Die Erfahrungen mit dieser Markierung sind, von ästhetischen und jederzeit korrigierbaren Kritiken abgesehen, positiv. Ein Grund kann in der homogenen und spezialisierten Nutzergruppe gesehen werden, die z.T. sogar in Spezialkursen auf die Nutzung von BISTRO vorbereitet wird. Ein weiterer Grund könnte darin gesehen werden, dass jeder in der tagtäglichen Konversation aktiv und passiv Kommunikationsstrukturen verwendet, so dass ein grammatikalischen Strukturen nachempfundener Aufbau auch bei komplexeren Datenansichten intuitiv erfassbar bleibt.[196]

Die Datenansichten sind zugleich die Arbeitsoberfläche für die Terminografen, womit nun nicht nur ein Feld in einem Eintrag bearbeitet werden kann, sondern alle Felder mit einem bestimmten Kriterium zugleich. Mit einem Befehl UPDATE können so alle Informationen, die sich auf ein bestimmtes Gesetz stützen, gleichzeitig geändert werden, wenn sich das Gesetz geändert hat. Durch die Normalisierung, d.h. die Eliminierung redundanter Daten in relationalen Datenbanken ist oft sogar nur noch die Änderung einer Tabelle nötig, auf die dann alle Datenansichten zugreifen. Terminografen können nun quer über alle Einträge und Datenkategorien Vergleiche anstellen.[197] Sogar die terminografischen Spezialwerkzeuge Termerkennung (*term recognition*, TR), Termextraktion

---

[193] Zum italienisch-deutschen Korpus siehe Gamper J., Construction of a Parallel Text Corpus Encoding Primary Data, in: Academia Nr: 18 (März - Juni 1999), EURAC, Bozen 1999, http://www.eurac.edu/Press/Academia/18/Art_13.asp : 30.3.2004. Zum italienisch-deutsch-ladinischen Parallelkorpus siehe Streiter O., Stuflesser M., Ties I. (2004), CLE, an aligned Trilingual Ladin-Italian-German Corpus. Corpus Design and Interface, S. 84-87 in: Carson-Berndsen J. (2004), Proceedings of the SALTMIL Workshop at the 4th International Conference on Language Resources and Evaluation (LREC) 2004, First Steps in Language Documentation for Minority Languages – Computational Linguistic Tools for Morphology, Lexicon and Corpus Compilation, European Language Resources Association (ELRA) Paris 2004, http://dev.eurac.edu:8080/autoren/publs/lrec9q.pdf : 30.3.2004.

[194] Neben den typischen Daten wie grammatikalischen Informationen sind z.B. auch das Rechtssystem, das Fachgebiet, die Bearbeitungsphase, die Empfehlung einer Benennung für die Normierung, der Normierungsstatus, die Alinierungsinformation für die Paralleltexte und verschiedene Metadaten als Tabellen implementiert.

[195] Streiter O. et al. (2004) a.a.O und konkret http://dev.eurac.edu:8080/cgi-bin/index/bistro?w=sqlselect&table=view_corp3&where=doc1+%3D+4464825&limit=120&order_by=p&xslttemplate=1 : 4.10.2004.

[196] Humphreys K. (1997), Formalising Pragmatic Information for Natural Language Generation, Dissertation University of Edinburgh, Centre for Cognitive Science, 1995. Die Online-Version von 1997 findet sich unter http://citeseer.ist.psu.edu/humphreys97formalising.html : 30.3.2004.

[197] Das geschieht technisch mit PROLOG-ähnlichen Befehlen. PROLOG steht für PROgrammieren mit LOGik (*Programming in Logic*) und wurde 1971 für die Verarbeitung natürlicher Sprache geschaffen und hat sich zu einer der gebräuchlichsten Sprachen in der Künstlichen Intelligenz entwickelt. Terminologen können mit PROLOG Ansichten komponieren, indem sie in bestimmte Knoten und Relationen zu einer virtuellen Hierarchie kombinieren. Damit kann jedes Eintragsmodell nachgeahmt oder komplexe deduktive Schlussfolgerungen erzeugt werden.

(*term extraction*, TE) und geordnete Fundstellen (*keyword-in-context*, KWIC) können in Datenansichten eingebaut werden und damit Teil des Nutzerdiskurses werden. Von den dynamisch erzeugten Dokumenten führen dynamisch erzeugte Verweise zurück zu den terminologischen Daten, zum Korpus oder wiederum zu den Spezialwerkzeugen, so dass die Darstellung gesammelten Wissens nahtlos in die Wissenssuche integriert ist. Damit wird die Termdarstellung ein Teil des Terminologiesammelns und das Terminologiesammeln ein Teil der individuellen Wissensrecherche.

## 8. Zusammenfassung

Das Ziel der Ablösung der traditionellen Terminografie durch das neue Paradigma dynamischer Termdarstellung mit aktiver (Motivierung zur Datenabfrage), interaktiver (Nutzerdialog mit Thema, Rhema und Fokus) und proaktiver (Einbeziehung der Nutzerwünsche und -ziele) Darstellungsstrategie scheint auch für laufende Terminografieprojekte erreichbar zu sein. Ein Mensch-Maschine-Diskurs von bekanntem zu noch unbekanntem terminologischem Wissen beruht auf dem Aufspalten traditioneller Einträge in elementare Informationseinheiten. Aus diesen Standardbausteinen kann man in unbegrenzter Kombinationsmöglichkeit Informationspakete schnüren, die wie für den individuellen Nutzer angefertigte terminologische Einträge aussehen. Die Abfolge der Informationspakete folgt einer Kommunikationsstrategie, die mit vier ineinander greifenden Grammatikmodulen implementiert werden kann. Jede Datenkategorie (Wortgrammatik), jedes Informationspaket (Satzgrammatik) und jede Abfolge von Informationspaketen (Textgrammatik) wird auf das Kommunikationsziel Wissensübermittlung ausgerichtet. Information sollte ausgeblendet werden, wo sie momentan störend oder überflüssig ist. Informationspakete müssen kohärent (widerspruchsfrei) und kohäsiv (zusammenhängend) verknüpft werden, wozu eine ‚Artikulationsgrammatik' die Verbindung des Gewussten mit dem Neuen herstellt. Solche Nutzerbedürfnisse werden bereits im elektronischen Handel, aber auch in verwandten Projekten der Wissensdarstellung erfolgreich berücksichtigt.

In der Terminologie steht dem menschlichen Nutzer ein sehr dichter Informationsspeicher gegenüber, so dass für die Rücksichtnahme auf Nutzerinteressen ein großer Spielraum besteht. Mit der hier vorgeschlagenen Methode kann darüber hinaus auch die Isolation terminologischen Wissens aufgehoben werden und eine Brücke geschlagen werden zur Termsammlung (keine Anpassung der Daten an ein Eintragsmodell), zur Termextraktion (vergl. Kapitel 4) sowie zur didaktischen Wissensvermittlung (vergl. Kapitel 5). Dynamisch dargestellte Terminologie wird dann weniger mit einem gedruckten Wörterbuch gemein haben als mit einer guten Lernsoftware.

Eine geeignete Technik für dynamische Termdarstellung steht zur Verfügung und wurde erfolgreich in BISTRO implementiert (20.000 Zugriffe externer Nutzer pro Monat). Erste Versuche mit der neuen Flexibilität zeigen, dass die grammatikalische Strukturierung grundsätzlich intuitiv nachvollziehbar ist. Die Nutzer treffen auf ein kommunikationswilliges Gegenüber, das ihnen individuell interessante Terminologie zusammenstellt und die individuelle Navigation unterstützt.

Produkt dynamischer Termdarstellung ist aber nicht nur die Veröffentlichung im Internet, sondern terminografische mono- und multidirektionale Papierwörterbücher für Fachleute, Übersetzer und Generalisten. Die Kreation einer neuen Datenansicht genügt, um Spezialglossare, *Translation Memories*, Definitionswörterbücher oder Listen mit Falschen Freunden ausgeben zu können. Die elementare Datenstruktur erlaubt außerdem effektiven Datentransfer und -rekombination. Bisher noch nicht ausgeschöpft wurde die Möglichkeit der Künstlichen Intelligenz durch Inferenz und Deduktion aus dem terminologischen Wissen.

Die hier vorgestellte Methode ist weithin auf wohlwollendes Interesse gestoßen und wird unter anderem Namen in verschiedene große Terminologieprojekte einfließen. Damit fördert sie den anstehenden Paradigmenwechsel in der Terminologie.

Während die Fachgebietserkennung (Kapitel 2) über die Terminografie hinaus von allgemeinem Interesse ist und dieses Kapitel für eine Neuorientierung in der Terminografie plädiert, soll nun in

Kapitel 4 ein spezifisch terminografisches Instrument zur Auffindung von Termkandidaten vorgestellt werden: Eine Termextraktion durch Beispielterme.

# Literaturangaben zu Kapitel 3

Apresjan J. D., Boguslavskij I. M., Iomdin L. L., Lazurskij A. V., Sannikov V. Z., Tsinman L. L. (1992), ETAP-2: The Linguistics of a Machine Translation System, S. 97-112 in: Meta, Bd. 37:1, http://www.erudit.org/revue/meta/1992/v37/n1/001895ar.pdf

Ardissono L., Goy A., Petrone G., Segnan M., Torasso P. (2003), Intrigue: Personalized Recommendation Of Tourist Attractions For Desktop And Handset Devices, S. 687-714 in: Applied Artificial Intelligence: Special Issue on Artificial Intelligence for Cultural Heritage and Digital Libraries (2003) 17:8-9, http://www.di.unito.it/~liliana/EC/aai03.pdf : 29.3.2004.

Beißwenger M., Storrer A., Runte M. (2003), Modellierung eines Terminologienetzes für das automatische Linking auf der Grundlage von WordNet, S. 95-104 in: Kunze C., Lemnitzer L., Wagner A. (Hrsgg.)(2003), Tagungsband des LDV-Forum 19 (2003): 1-2, Sonderheft mit Beiträgen des GermaNet Workshops zu Anwendungen des deutschen Wortnetzes in Theorie und Praxis.

Bourigault D., Jacquemin C., L'Homme M.-C., (Hrsgg.) (2001), Recent Advances in Computational Terminology, John Benjamins Publishing Company Amsterdam 2001.

Brinkmann K.-H., Schulz J., Tanke E., (1969), Das Wörterbuch aus der Maschine, S. 9-15 in: Data Report 4/4 1969; nachgedruckt in: Graham J. D., Grewe K., Reisen U. (Hrsgg.) (1995), Terminologiearbeit. Theorie und Praxis, in: Festschrift für Eberhard Tanke zum 75. Geburtstag, Deutscher Terminologie-Tag Köln 1995.

Budin G. (2002), Der Zugang zu mehrsprachigen terminologischen Ressourcen – Probleme und Lösungsmöglichkeiten, in: In Mayer F., Schmitz K.-D., Zeumer J. (Hrsgg.), eTerminologie - Professionelle Terminologiearbeit im Zeitalter des Internet, Tagungsakte des Symposions “eTerminology”, Deutscher Terminologie-Tag (DTT) Köln 2002.

Bullo F., Ciola B., Coluccia S., Maganzi Gioeni D'Angiò F., Mayer F., Treiber A., Voltmer L. (2003), Terminologisches Wörterbuch zum Vertragsrecht: italienisch/deutsch Dizionario terminologico del diritto dei contratti italiano – tedesco, C.H.Beck München, Athesia Bozen, Stämpfli Bern, Linde Wien 2003.

De Carolis B., De Rosis F., Grasso F., Rossiello A., Berry D. C., Gillie T. (1996), Generating recipient-centered explanations about drug prescription, S. 123-145 in: Artificial Intelligence in Medicine, Vol. 8 (1996): 2.

De Carolis B. (1999), Generating Mixed-Initiative Hypertexts: A Reactive Approach, S. 71-78 in: Proceedings of the 1999 International Conference on Intelligent User Interfaces (IUI ’99), IUI San Diego 1999, http://citeseer.ist.psu.edu/decarolis99generating.html :29.3.2004.

De Carolis B., Pizzutilo S., Palmisano I. (2003), D-ME: Personal Interaction in Smart Environments, S. 388-392 in: Lecture Notes in Computer Science, Vol. 2702, Springer-Verlag Heidelberg 2003.

Firth J. R., Palmer F. (Hrsgg.) (1968), Selected Papers of J.R. Firth 1952-59, Longman's Linguistic Library London 1968.

Gamper J. (1999), Construction of a Parallel Text Corpus Encoding Primary Data, in: Academia Nr: 18 (März - Juni 1999), EURAC, Bozen 1999, http://www.eurac.edu/Press/Academia/18/Art_13.asp :30.3.2004.

Giese H., Kleppin K., Schlagwort ‚Falsche Freunde', S. 204 in: Glück H (2000), Metzler Lexikon Sprache, 2. Auflage J. B. Metzler Verlag Stuttgart Weimar 2000.

Graham J. D., Grewe K., Reisen U. (Hrsgg.) (1995), Terminologiearbeit. Theorie und Praxis, in: Festschrift für Eberhard Tanke zum 75. Geburtstag, Deutscher Terminologie-Tag Köln 1995.

Halliday M. A. K. (1994), An Introduction to Functional Grammar, 2. Aufl. Edward Arnold London 1994.

Holmes-Higgin P., and Khurshid A. (1996), Is your Terminology in Safe Hands? Data Analysis, Data Modelling and Term Banks, Terminology and Knowledge Engineering, S. 215-224 in: Proceedings of the 4th International Congress on Terminology and Knowledge Engineering Vienna (TKE '96), INDEKS-Verlag Frankfurt 1996.

Hovy E. H. (1993), Automated Discourse Generation Using Discourse Structure Relations, S. 341-385 in: Artificial Intelligence (AI) 63 (1993): 1-2,
http://citeseer.ist.psu.edu/hovy93automated.html :29.3.2004.

Humphreys K. (1997), Formalising Pragmatic Information for Natural Language Generation, Dissertation University of Edinburgh, Centre for Cognitive Science, 1995.
http://citeseer.ist.psu.edu/humphreys97formalising.html :30.3.2004.

Malinowski, B. (1923), The Problem of Meaning in Primitive Languages, Supplement I auf S. 296-336 in: Ogden C. K., Richards I. A. (Hrsgg.), The Meaning of Meaning, 8. Auflage Harcourt Brace & World New York 1946.

Mann W., Thompson, S. (1988), Rhetorical Structure Theory: toward a functional theory of text organization, S. 243-281 in: Text 8 (1988): 3.

Matthiessen C., Bateman J. A. (1991), Text generation an Systemic-Functional Linguistics - Experiences from English and Japanese, Pinter Publishers London 1991.

Mayer F. (1998), Eintragsmodelle für terminologische Datenbanken: ein Beitrag zur übersetzungsorientierten Terminographie, Forum für Fachsprachen-Forschung Bd. 44, Narr Tübingen 1998.

Mayer F. (1996), The representation of inconsistent relationships in termbanks, S. 225-232 in: Galinski C., Schmitz K.-D. (Hrsgg.) (1996), Terminology and Knowledge Engineering Conference 1996 (TKE ’96), Indeks Frankfurt a. M. 1996.

Mel’čuk I. (2001), Communicative Organization in Natural Language: The Semantic - Communicative Structure of Sentences, Studies in Language Companion Series 57, John Benjamins Publishing Company Amsterdam 2001.

Melby A. K., Wright S. A. (1999), Leveraging terminological data for use in conjunction with lexicographical resources, S. 544-569 in: Proceedings of the Fifth International Congress on Terminology and Knowledge Engineering Conference Innsbruck (TKE ’99), TermNet Innsbruck 1999,
http://www.ttt.org/TKE-99.pdf :25.3.2004.

Sager J. C. (1990), A practical Course in Terminology Processing, John Benjamins Publishing Company Amsterdam 1990.

Schmidt-Wigger A. (1998), Building consistent terminologies, Poster in: Proceedings of the First Workshop on Computational Terminology (COMPUTERM'98) at the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics (COLING-ACL'98), ACL University of Montreal (Hrsgg.) Montreal 1998,
http://www.iai.uni-sb.de/docs/term2.pdf : 1.10.2004.

Schmitz K.-D. (2002), Towards a uniform environment for representing terminologies within ISO, Präsentation auf der Terminology and Knowledge Engineering (TKE) Konferenz Nancy 2002,
http://tke2002.loria.fr/Doc/workshops/ws2/ws2_kds.ppt :18.3.2004.

Streiter O., Stuflesser M., Ties I., (2004), CLE, an aligned Trilingual Ladin-Italian-German Corpus. Corpus Design and Interface, S. 84-87 in: Carson-Berndsen J. (2004), Proceedings of the SALTMIL Workshop at LREC 2004, First Steps in Language Documentation for Minority Languages – Computational Linguistic Tools for Morphology, Lexicon and Corpus Compilation, European Language Resources Association (ELRA) Paris 2004,
http://dev.eurac.edu:8080/autoren/publs/lrec9q.pdf : 30.3.2004.

Streiter O., Voltmer L. (2003), A Model for Dynamic Term Presentation, S. 201-204 in: Tagungsband der TIA-2003 Konferenz, LIIA – ENSAIS, Université Marc Bloch (Hrsgg.) Strasbourg 2003, http://dev.eurac.edu:8080/autoren/publs/ModDynTerm.pdf.gz: 13.9.2004.

Internetquellen in Zitierreihenfolge:
http://www.uni-leipzig.de/~xlatio/software/soft-termiman.htm : 4.10.2004.
http://en.wikipedia.org/wiki/List_of_false_cognates : 4.10.2004.
http://www.iim.fh-koeln.de/webterm/ : 4.10.2004
http://www.w3.org/TR/wbxml/ : 4.10.2004.
http://mtt2003.linguist.jussieu.fr/ : 4.10.2004.
http://www.w3.org/WAI : 4.10.2004.
http://www.cs.vassar.edu/CES/ : 4.10.2004.
http://www.xml-ces.org/http://www.cs.vassar.edu/XCES : 4.10.2004.
http://www.w3.org/XML/ : 4.10.2004.
http://www.w3.org/Style/CSS/ : 4.10.2004.
http://www.w3.org/TR/REC-CSS2/aural.html : 4.10.2004.
http://www.w3.org/XSLT : 4.10.2004.
http://www.w3.org/TR/xsl/slice6.html : 4.10.2004.
http://www.mip.berkeley.edu/mip/related/thesaurus/thesaurus.pdf: 4.10.2004.
http://www.rpbourret.com/xml/XMLAndDatabases.htm : 4.10.2004.
http://www.dynamicmultimedia.com.au/peba/system.html : 4.10.2004.
http://www.cmis.csiro.au/iit/Projects/DDD/ : 4.10.2004.
http://www.iuiconf.org/pastiui.html : 4.10.2004.
http://privacy.net/analyze/ : 4.10.2004.
http://atilf.atilf.fr : 4.4.2005.
http://www.hytex.info/ : 4.10.2004
http://dev.eurac.edu:8080/cgi-bin/index/bistro?w=sqlselect&table=view_corp3&where=doc1+%3D+4464825&limit=120&order_by=p&xslttemplate=1 : 4.10.2004.

Kapitel 4

# IV. Termextraktion durch Beispielterme

## Automatische Erkennung von Rechtsbegriffen durch Beispielbenennungen zum Terminologieaufbau und zur Indizierung

In diesem Kapitel werden zunächst die drei Ansätze der Termextraktion (TE) vorgestellt (1.), und zwar die linguistische oder regelbasierte TE (2.), die statistische TE, bei der auch auf verschiedene Assoziationsmaße eingegangen wird (3.), und schließlich die TE durch Beispielterme (4.). Anschließend werden die Ansätze miteinander verglichen (Punkt 5.) und dargelegt, wie eine Termextraktion auch zur Indizierung von Rechtstexten verwendet werden kann (6.). In Punkt 7. werden die statistischen Vergleichsmaße des IR vorgestellt, mit denen die Ergebnisse von TE gemessen werden können. Der Punkt 8. stellt dann eigene Experimente vor, die in Punkt 9. mit den Ergebnissen in der Literatur verglichen werden, um in 10. zu einer Bewertung der eigenen Experimente zu gelangen. Zuletzt werden noch Überlegungen zur Übertragung der Ergebnisse auf die Textindexierung angestellt (11.).

### 1. Einführung in die drei Termextraktionsmethoden

Terminologie und Terminografie beschäftigen sich mit Termen[198]. Terme werden in Texten oder Kontexten verwendet. Es ist zeitaufwendig und mühsam, Term für Term aus dem Text oder Kontext zu lösen. Daher versucht die Computerlinguistik, Terme automatisch aus elektronischen Texten herauszusuchen.

Die computergestützte Termextraktion (*computer aided terminology extraction*) ist ein neueres Forschungsgebiet in der maschinellen Sprachverarbeitung (*natural language processing* – NLP).

Termextraktion exzerpiert Terminologie aus Texten. Die computergestützte Termextraktion analysiert einen maschinenlesbaren Text und filtert die darin enthaltenen Termkandidaten mit Hilfe des Computers heraus. Termkandidaten sind Wörter oder Phrasen, wie sie in Glossaren oder Wörterbüchern stehen oder wie sie als Schlagwörter und für Indizes verwendet werden.

Die Termextraktion wird im Wesentlichen für drei **Zwecke** gebraucht. Man kann

1. einen Index, Schlagwörter o.ä. erzeugen (*indexing*),
2. bereits bekannte Terme in Texten wieder finden (Terminologieerkennung - *term recognition*), um die Texte zu annotieren oder klassifizieren oder
3. neues Wissen automatisch erzeugen (*automatic knowledge acquisition*).

Die automatische Wissenserzeugung durch Termextraktion ist meist das Finden noch nicht beschriebener Terminologie (*term discovery*). Sollen die gefundenen Terme einen Terminologiebestand erweitern, spricht man von *term enrichment*, soll eine Terminologie ganz neu aufgebaut werden von *term acquisition*.

[198] Terme sind ‚Termini' oder ‚Benennungen', vergl. Fußn. 24.

Die Begriffe, mit denen sich die Terminologie und Terminografie beschäftigt, werden als bereits gegeben vorausgesetzt und die Termextraktion beschränkt sich auf das Auffinden und den expliziten Erwerb dieser Terme, es werden aber keine Terme neu geschaffen.[199] Das Bestimmen von Schlagwörtern und Indizes zum Suchen, Wiederfinden und Einteilen von Dokumenten kann man hingegen als kreativen Prozess betrachten, weil Elemente verwendet werden können, die außerhalb des Indexes keine Bedeutung haben.[200] Hier sind die Terme nur Mittel, während sie bei der Terminologieerkennung bereits selbst das Ziel darstellen. Eine Termextraktion zur Indizierung kann daher schlechtestenfalls unzweckmäßig oder nicht zielführend sein, während eine Terminologieerkennung fehlerhaft sein kann. Eine Folge davon ist, dass fast ausschließlich bei der Terminologieerkennung von Hand nachbearbeitet und verbessert wird, und zwar sehr häufig.

Termextraktion wird in **zwei computerlinguistische Teilaufgaben** zerlegt. a) Einerseits muss entschieden werden, welche Teile eines Textes zusammengehören (*unithood problem*), b) andererseits muss erkannt werden, welche zusammengehörenden Teile tatsächlich ein Term sind (*termhood problem*). Ein Adjektiv gehört stets zu seinem Substantiv (Bsp.: indirekte Steuer), aber nicht jeder Kombination aus Adjektiv und Substantiv wird Termwert zugesprochen (nicht: hohe Steuer). Oft wird ein Text zuerst in Einheiten zerlegt, die dann auf ihre Termqualität untersucht werden.

Es gibt folgende Ansätze zur Termextraktion:

- linguistische Ansätze
- statistische Ansätze und
- beispielbasierte Ansätze.[201]

Während die linguistischen Ansätze sprachliches Wissen (z.B. morphologische oder syntaktische Informationen) einsetzen, versuchen statistische Methoden über den Vergleich vieler Textbruchstücke (z.B. nach Häufigkeitsverteilung, Assoziationskoeffizienten oder mit statistischen Testverfahren) zu Erkenntnissen über die Struktur der Texte und die Lage von Termen zu gelangen. Ersteres erfordert tiefgehendes Wissen über die Sprache, letzteres eine breite Untersuchungsgrundlage. Heute werden diese beiden Ansätze meistens zu einem hybriden Ansatz verbunden.[202]

### 2. Linguistische Termextraktion

Linguistische Ansätze verwenden morphologische, syntaktische oder semantische Informationen aus sprachspezifischen Anwendungen. Ihr Hauptziel ist das Erkennen von Spracheinheiten. Das Sprachmodul soll die Zusammensetzung von Termen besonders effizient und genau analysieren. Das sprachliche Wissen bezieht sich beispielsweise auf die Anzahl der Wörter eines Terms, auf besondere Vor- oder Nachsilben und auf grammatikalische Anpassungen des Terms (Mehrzahl, Konjugation). Diese Analyse erledigen morphologische Analyseprogramme, Part-of-Speech-Tagger und

---

[199] Zur Einteilung der Termextraktion siehe Zielinski D. (2002), Computergestützte Termextraktion aus technischen Texten, Diplomarbeit Universität des Saarlands, Saarbrücken 2002, http://www.iai.uni-sb.de/~mt-dept/texte/zielinski.pdf : 4.10.2004.

[200] Ein einfaches Beispiel sind Wortwurzeln, die in Texten nur mit Endungen vorkommen, aber bei der Suche hilfreich sein können.

[201] Eine Einteilung in linguistische, statistische und hybrid linguistisch-statistische Ansätze findet sich bei Cabré Castellvì M.T., Estopà Bagot R., Palatresi J. V., (2001), Automatic term detection: A review of current systems, S. 53-89 in: Bourgault D., Jacquemin C., L'Homme M. C. (Hrsgg.) (2001), Recent Advances in Computational Terminology, Natural Language Processing Band 2, John Benjamins Publishing Company, Amsterdam/ Philadelphia 2001.

[202] Maynard D., Ananiadou S. (2001), Term extraction using a similarity-based approach, S. 53-89 in: Bourigault D., Jacquemin C., L'Homme M. C. (Hrsgg.) (2001) a.a.O., http://citeseer.nj.nec.com/maynard99term.html : 4.10.2004.

Parser.[203] Dazu werden z.T. auch Stoppwortlisten verwendet, in denen Wörter stehen, die nie in einer bestimmten Position (Termanfang, letztes Wort des Terms oder Mittelstellung) eines Terms auftauchen.

## 3. Statistische Termextraktion

Statistische Ansätze der Termextraktion beruhen auf der Annahme, dass Terme aus lexikalischen Einheiten bestehen, die statistisch signifikant öfter gemeinsam auftauchen, als bei der Kombination unabhängiger Einheiten zu erwarten wäre. So können Mehrworttermkandidaten gefunden werden. Im Folgenden werden die wichtigsten Methoden kurz vorgestellt.

Die Häufigkeit des gemeinsamen Auftretens von lexikalischen Einheiten ist noch kein alleiniger Garant für deren Termeigenschaft, weil häufig nebeneinander stehende Funktionswörter keine Termkandidaten sind. Man verschärft daher die Annahme und sucht nicht nur nach häufigen Begriffen, sondern nach Fachbegriffen. Man nimmt an, dass Fachbegriffe eines Dokuments in diesem häufiger verwendet werden als in anderen Dokumenten. Man sucht also jene Kombination bestimmter lexikalischer Einheiten (Fachwörter) oder morphosyntaktischer Konstruktionen (Fachjargon), deren Frequenz in wenigen Dokumenten hoch, in allen Texten hingegen niedrig ist.

### 3.1. TF.IDF

Mathematisch drückt dies die Standardformel TF.IDF aus dem *Information Retrieval* aus. In dieser Formel wird die Häufigkeit eines Terms (TF= *term frequency*) in einem Dokument durch die Häufigkeit dieses Terms in allen Dokumenten geteilt, bzw. mit der *inverted document frequency* multipliziert:

$$TF.IDF_x = \frac{TCF_x}{DF_x}$$

mit TC = Termkandidat, D = Dokument,; $F_x$ = Häufigkeit (*frequency*) von x.

### 3.2. *weirdness-ratio*

Dieselbe Idee steht auch hinter der *weirdness ratio*[204], in der relative Häufigkeiten für TF und IDF verwendet werden:

$$weirdness\ ratio_x = \frac{\frac{TCF_x}{\#\{TC\}}}{\frac{DF_x}{\sum_{d=1}^{d=m} doc_j}}$$

Häufigkeitszählungen eignen sich gut für die Berechnung ununterbrochener Zeichenketten und feststehender Phrasen. Wenn die lexikalischen Einheiten jedoch nicht unmittelbar nebeneinander stehen (z.B. wegen Partikelverben oder Einschüben in die Phrase), dann muss die Beziehung zwischen Wörtern in einem Satz bestimmt werden. Eine Möglichkeit, die Assoziation von lexikalischen Einheiten zu messen, ist die Berechnung ihres Abstandes im Text. Als Abstandsmaße werden der

[203] Ein Part-of-Speech-Tagger ist ein Werkzeug zur automatischen Auszeichnung von Wörtern nach ihrer Wortart, während ein Parser ein Werkzeug zur automatischen Analyse insbesondere der grammatischen Struktur von sprachlichen Ausdrücken ist. Die Wirkungsweise dieser Instrumente ist am leichtesten durch Beispiele zu demonstrieren, etwa über die Links unter http://www.ifi.unizh.ch/CL/InteractiveCLtools/index.php : 4.10.2004.

[204] Brekke M., Myking J., Ahmad K. (1996), Terminology management and lesser-used living languages: A critique of the corpus-based approach, S. 179–189 in: Sandrini P. (Hrsg.) (1996), Proceedings of the 4th International Congress on Terminology and Knowledge Engineering (TKE'96) Innsbruck, TermNet Wien 1996.

statistische Mittelwert (*mean*), die Varianz (*variance*) oder die Standardabweichung (*deviation*) verwendet.[205]

Verfahren, die sich nach der Häufigkeit richten, funktionieren gut für Einwortterme, sie können aber nicht maßstäblich auf Zwei- und Dreiwortterme vergrößert werden. Wenn eine derartige Termextraktionsmethode für 10.000 Wörter gut funktioniert, dann würde die Erweiterung auf Zweiwortausdrücke 100.000.000 Wörter benötigen, um gleich gute Ergebnisse aufzuweisen. Für Termkandidaten aus drei Wörtern wären $10.000^3$ = 1.000.000.000.000 Wörter nötig. Dieser Zusammenhang ist als *sparse-data problem*[206] bekannt und findet sich in allen statistischen Maßen, die die Häufigkeit eines Termkandidaten verwenden.

Die Häufigkeit eines Termkandidaten ist aber nur ein Referenzmaß.[207] Man kann auch die Korrelation der lexikalischen Einheiten A und B eines Termkandidaten untereinander messen. Dazu trägt man in eine Tabelle ein, wie häufig man die vier logisch möglichen Fälle [A; B], [A; nicht-B], [nicht-A; B] und [nicht-A; nicht-B] auftreten:

**Tabelle 7: Kontingenztabelle der beobachteten Ereignisse**

| | Wort2 = B | Wort2 ≠ B | ∑ |
|---|---|---|---|
| Wort1 = A | $O_{11}$ | $O_{12}$ | $R_1$ |
| Wort1 ≠ A | $O_{21}$ | $O_{22}$ | $R_2$ |
| ∑ | $C_1$ | $C_2$ | N |

O ist *occurrence*/Auftreten. Die Zahl 1 im Index bedeutet „tritt auf", die Zahl 2 bedeutet „tritt nicht auf". $C_1$ ist die Häufigkeit von [B], $C_2$ die von [nicht-B], $R_1$ die von [A] und $R_2$ die von [nicht-A]. N ist die Summe aller Dokumente mit $C_1+C_2=R_1+R_2$.

Aus der Häufigkeit von A und B lässt sich eine Erwartung (*expectancy*) für das Zusammentreffen von A und B in einem Dokument berechnen. Ebenso lassen sich die Wahrscheinlichkeiten für [A; nicht-B], [nicht-A; B] und [nicht-A; nicht-B] aus den Summen der beobachteten Häufigkeiten schätzen. Die entstehende Wahrscheinlichkeitstabelle heißt Kontingenztabelle des Wortpaars [A; B]:

**Tabelle 8: Kontingenztabelle der erwarteten Ereignisse**

| | $Wort_2$ = B | $Wort_2$ ≠ B |
|---|---|---|
| $Wort_1$ = A | $E_{11} = \frac{R_1 \cdot C_1}{N}$ | $E_{12} = \frac{R_1 \cdot C_2}{N}$ |
| $Wort_1$ ≠ A | $E_{21} = \frac{R_2 \cdot C_1}{N}$ | $E_{22} = \frac{R_2 \cdot C_2}{N}$ |

E ist die Erwartung (*expectance*). R, C und N sind die beobachteten Häufigkeiten der Kontingenztabelle 1. Die Zahl 1 im Index bedeutet „tritt auf" die Zahl 2 bedeutet „tritt nicht auf".

[205] Das Abstandsmaß geht auf Smadja, F. (1993) zurück: Retrieving collocations from text: Xtract, S. 143-177 in: Computational Linguistics 19(1) mit weiteren Nachweisen, http://acl.ldc.upenn.edu/J/J93/J93-1007.pdf: 4.10.2004. Zur Berechnung ausführlicher Zielinski D., 2002, a.a.O.

[206] Problem der geringen Häufigkeiten oder der erforderlichen Datenmenge.

[207] Die meisten Referenzmaße sind im Perl Modul *N-gram Statistics Package* (letzte Version 0.71 vom 17.6.04) implementiert, das unter http://www.d.umn.edu/~tpederse/nsp.html :13.9.2004 kostenlos heruntergeladen werden kann.

Je weiter der tatsächlich gemessene Wert vom errechneten Wert abweicht, umso außergewöhnlicher ist das Ereignis und umso kleiner ist die Wahrscheinlichkeit, dass es sich um Zufall handelt. Es kann sich um eine positive oder negative Abweichung handeln.

3.3. *mutual information*

Es gibt viele verschiedene Maße, die Abweichung zwischen den tatsächlichen (Tabelle 7) und den erwarteten Ereignissen (Tabelle 8) zu berechnen. Ein in der Korpuslinguistik häufig gebrauchtes Maß ist die *mutual information* (MI):

$$MI = \frac{O_{11}}{E_{11}}$$

Die MI ist also das beobachtete gemeinsame Auftreten von A und B (also das Feld $O_{11}$), dividiert durch die statistische Wahrscheinlichkeit gemeinsamen Auftretens ($E_{11}$). Den Wert $O_{11}$ erhält man durch Auszählen, der Wert von $E_{11}$ ist das Produkt der Häufigkeit von A mit der Häufigkeit von B, geteilt durch die Anzahl der Dokumente.

Das bedeutet erstens, dass nur zwei der acht informationstragenden Felder miteinander verglichen werden. Zweitens wird keine Wahrscheinlichkeit mit einem Wert zwischen Null und Eins berechnet, sondern ein Wert auf einer nicht linearen Skala von Null bis Unendlich. Die MI ist allerdings einfach zu berechnen, kann auf jeden Text angewandt werden und kommt ohne linguistische Informationen oder Termdatenbanken aus.

Dadurch kann man mit MI-Werten nur Aussagen über höhere oder niedrigere Wahrscheinlichkeit machen, man kann diese Wahrscheinlichkeit aber nicht quantifizieren oder mit ihr rechnen. Tatsächlich ist die MI vor allem bei kleinen Häufigkeiten ein ungeeignetes Maß.[208] Außerdem ordnen Häufigkeitsmaße nur Termkandidaten mit gleicher Wortzahl korrekt. Sobald aber Einwort- und Mehrwortterme gegeneinander konkurrieren, werden Birnen mit Äpfeln verglichen und eine Gruppe wird benachteiligt. Viele Ähnlichkeitsmaße können mehrere Wörter überhaupt nicht zueinander in Beziehung setzen, weil die Berechnungsformel nicht definiert ist. Selbst das relativ einfache MI-Maß erforderte zwei dreidimensionale Kontingenztabellen, die weder statistisch noch informationstechnisch definiert sind.

3.4. Weitere Assoziationsmaße

Andere Assoziationsmaße (*association measures*) sind das $\chi^2$-Maß (=*chi2*-Maß = *chi-square* = *chi*-Quadrat), der *t-score* und die *likelihood ratio*.

$$\chi^2 = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

O und E beziehen sich auf obige Kontingenztabelle. i und j können die Werte 1 oder 2 annehmen. Es gibt daher vier $\chi^2$-Formeln.

Das *chi2*-Maß nimmt keine normalverteilten Wahrscheinlichkeiten an. Es eignet sich vor allem zur Untersuchung von häufigen linguistischen Phänomenen und bei großen Textmengen. Allerdings gibt *chi2* nur an, ob ein Zusammenhang besteht, aber nicht wie stark der Zusammenhang ist und ob er negativ oder positiv ist. Daher wird es häufig in Formeln integriert, die solche Aussagen zulassen.

Der *t-score*

$$t - score = \frac{O_{11} - E_{11}}{\sqrt{O_{11}}}$$

und die *log-likelihood ratio*

[208] Vergl. Zielinski D. (2002) a.a.O.

$$Log\text{–}likelihood = 2\sum_{ij} O_{ij} \cdot \log_2\left(\frac{O_{ij}}{E_{ij}}\right)$$

sind zwar besser für kleine Häufigkeiten geeignet, aber letztere ist nicht definiert für den Fall, dass eines der Wörter eines Termkandidaten nicht zumindest auch einmal alleine auftritt.[209]

Insgesamt kann man festhalten, dass alle Häufigkeitsmaße Termen einen nummerischen Wert zuweisen, nach dem die Termkandidaten geordnet werden. Ein willkürlich gewählter Schwellenwert trennt die Termkandidaten von den Nichttermen. Die Zusammengehörigkeit von Wörtern wird nur oberflächlich untersucht. Zum einen werden nur Wortsequenzen von vorher bestimmter Länge untersucht (z.B. Zweiwortterme), zum anderen werden Phrasengrenzen nicht beachtet, so dass zu einem Term gehörende Wörter unterschlagen oder nicht zu einem Term gehörende Wörter hinzugefügt werden.

Daher versucht ein anderer statistischer Ansatz, die Grenzen der Termkandidaten zu erkennen. Wenn man die Grenze einer Nominalphrase als zwischen einem Wort und dem ersten Wort eines Termkandidaten sowie zwischen dem letzten Wort eines Termkandidaten und dem nachfolgenden Wort festlegt, dann können Termkandidaten beliebige Länge haben. Damit vermeidet man das *sparse-data* Problem, denn man kann auch in kurzen Texten lange Termkandidaten finden. Wenn man Termanfang und Termende aufeinander bezieht, also den gesamten Termkandidaten als Grenze definiert, dann hat man wieder das *sparse-data* Problem. Um die Grenze herauszufinden, kann man kann im Grunde jedes Ähnlichkeitsmaß verwenden. Zwischen den schwach korrelierenden Wörtern liegen dann die Phrasengrenze. Oft findet man die Grenzen eines Terms aber durch so genannte Entropiemaße.

Die Entropie (*entropy*) ist der Grad der Unordnung bzw. des Zufalls und dient als Maß für die Schwierigkeit der Vorhersage eines Ereignisses. Eine Definition dieses Maßes ist die durchschnittliche Bitlänge zur Beschreibung dieses Ereignisses. Wenn alle möglichen Ereignisse *n* gleich wahrscheinlich sind, dann muss zur Beschreibung des tatsächlichen Ereignisses die entsprechende Zahl zwischen 1 und *n* berichtet werden. Dazu benötigt man log (*n*) Bits. Wenn nicht alle Ereignisse gleich wahrscheinlich sind, dann wählt man für die häufiger auftretenden Ereignisse eine kleinere Bitfolge und für die selteneren eine längere. Dadurch kann man Ereignisse mit der Wahrscheinlichkeit *p* (*probability*) theoretisch durch $-\log_2 p$ Bits beschreiben.

Je größer die Entropie[210] zwischen lexikalischen Einheiten ist, umso schwieriger ist die Voraussage der nächsten lexikalischen Einheit und umso unwahrscheinlicher ist die Zusammengehörigkeit dieser beiden Einheiten.

## 4. Termextraktion durch Beispielterme

### 4.1. Die fünf Phasen der Termextraktion durch Beispielterme

Der **beispielbasierte Ansatz** in der Verarbeitung natürlicher Sprache (*natural language processing* = NLP) zeichnet sich dadurch aus, dass das Trainingsmaterial von der gleichen Art ist wie das Ergebnis. Es gibt beispielsweise Programme, die mit Syntaxbäumen gefüttert werden, um das Erstellen von Syntaxbäumen zu lernen. Seit 1981 gibt es die Idee zur beispielbasierten maschinellen Übersetzung, bei der Übersetzungen eingegeben werden, um Übersetzungen zu finden.[211] Der Vor-

[209] Daille B. (1995), Combined approach for terminology extraction: lexical statistics and linguistic filtering, S. 515-521 in: Proceedings of the 15th International Conference on Computational Linguistics Kyoto (COLING 94) 1994, ICCL Kyoto 1994, http://acl.ldc.upenn.edu/C/C94/C94-1084.pdf : 4.10.2004.

[210] Entropie bezeichnet in der Informationstheorie den mittleren Informationsgehalt eines Nachrichtensymbols in Bit.

[211] Somers H. (2003), Machine Translation: Latest Developments, S. 512-528 in: Mitkov R. (Hrsg.) (2003), The Oxford Handbook of computational Linguistics, Oxford University Press Oxford 2003. Beispielbasierte Übersetzung ist nicht

teil beispielbasierter Ansätze gegenüber regelbasierten Ansätzen ist, dass keine abstrakten Regeln erstellt werden müssen. Die als Trainingsmaterial notwendigen Beispiele kann man also durch Auswahl aus existierenden Beispielen manuell oder automatisch erzeugen oder auch erfinden. Es ist keine komplexe Formalisierung des sprachlichen Wissens nötig. Beispiele zu Regeln und Ausnahmen können nebeneinander, also nicht-hierarchisch aufgelistet werden. Außerdem funktioniert der beispielbasierte Ansatz bereits mit wenigen Beispielen, im Extremfall mit einem einzigen. Jedes weitere gefundene Beispiel kann dem Trainingsmaterial hinzugefügt werden, wodurch das System weiter lernt und sich die Leistung weiter verbessert.

Die Termextraktion durch Beispiele verläuft in fünf Phasen:

1. Beispiele auswählen
2. maschinenlesbar beschreiben, wie aus Beispielen Muster zu erzeugen sind (Formalisierung der Musterart)
3. aus den Beispielen termtypische Muster erzeugen
4. mit Hilfe der Muster die Texte durchsuchen und Termkandidaten finden
5. Ordnung der Termkandidaten nach ihrer Termwahrscheinlichkeit

Eine Besonderheit der beispielbasierten Termextraktion ist, dass die in Punkt 1 beschriebene Aufteilung der Arbeitsaufgabe in die zwei Probleme *termhood* und *unithood* sich in den Arbeitsschritten widerspiegelt: Die Phasen 1 bis 4 versuchen das *unithood*-Problem zu lösen, erst die 5. Phase geht das *termhood*-Problem an.

Bei der **Termextraktion durch Beispielterme** werden vorhandene Terme eingegeben, um als Extraktionsergebnis Termkandidaten zu erhalten. Die zugrunde liegende Annahme ist, dass Kombinationen lexikalischer Einheiten, die den Beispieltermen hinreichend ähnlich sind, zusammengehören und selbst einen Term bilden. Wenn die Trainingsterme aus einer Termdatenbank stammen, dann werden die extrahierten Terme der expliziten oder impliziten Termdefinition dieser Datenbank entsprechen. Enthält die Termdatenbank nur Nominalphrasen, dann werden nur Nominalphrasen extrahiert. Sind auch Verbalphrasen unter den Beispieltermen, dann werden auch die extrahierten Terme dieselbe Zusammensetzung haben. Wenn man noch keine traditionell erstellten Terme hat und stattdessen Wörterbucheinträge verwendet, dann werden Terme extrahiert, die zu diesem Wörterbuch passen. Wenn das nicht gewünscht ist, dann sollte man den Teil des Wörterbuchs als Beispielterme verwenden, der der gewählten Termdefinition entspricht.

### 4.2. Formalisierungsparameter der TE mit Beispieltermen

Beispielbasierte Ansätze bei der Verarbeitung natürlicher Sprachen benutzen vorgegebene Lösungen zum Auffinden ähnlicher Lösungen in noch unbearbeiteter Umgebung. Es müssen also nicht nur Lösungen, sondern auch Ähnlichkeitsparameter eingegeben werden. Im Gegensatz zu regelbasierten Ansätzen müssen bei beispielbasierten Ansätzen zwar keine Regeln vorgegeben werden, aber doch Parameter, nach denen Regeln erzeugt werden sollen. Der Mensch muss also die Hypothesen darüber, worin sich alle Beispiele ähnlich sind, formalisieren.

Bei natürlichen Sprachen würde man z.B. vermuten, dass Terme gegenüber dem restlichen Test Besonderheiten aufweisen bezüglich ihrer Unterteilung in Wörter, ihrer Wortlänge, Groß- und Kleinbuchstaben, der Häufigkeit von Vor- und Nachsilben und der Verwendung unterschiedlicher Buchstaben. Schwieriger zu instrumentalisieren sind die Variablen der Kodierung für eine Bildschrift, Zeichenschrift oder vor allem für binäre Information.

---

gleichbedeutend mit *Translation Memory* (TM), weil bei ersterer eine fertige Übersetzung ausgegeben wird, beim TM nur evtl. Vorschläge zu möglichen Übersetzungen, über deren Verwendung der Übersetzer selbst entscheidet.

Ein Beispiel für die Formalisierung eines graphischen Musters ist die Regel klein-klein-klein. Dieses **Groß- und Kleinschreibungsmuster** besagt, dass lexikalische Einheiten dann ähnlich sind, wenn drei aufeinander folgende Wörter am Wortanfang kleingeschrieben sind. Dieses Muster wurde von dem ladinischen Beispielterm *tofla de comun* erzeugt und wurde für die Extraktion ladinischer Terme verwendet.

Die Erfüllung eines graphischen Musters allein wäre natürlich nicht strikt genug. Muster anderer Art müssen hinzugefügt werden. Der Beispielterm *tofla de comun* erzeugt z.B. das **Affixmuster** **a-de-*n*. Dieses Muster passt auf alle Kombinationen von drei Wörtern („-“ steht für ein Leerzeichen), in denen das erste Wort auf a endet („*“ steht für beliebige Buchstaben), dann das Wort „de“ kommt und dann ein auf n endendes Wort. Es passt daher auch auf die ladinischen[212] Terme *ćiasa de comun* und *contlamada de comun*.

Bei den beiden Musterbeispielen fällt auf, dass es im graphischen Muster der Groß- und Kleinschreibung sehr viel weniger Varianz gibt als im Affixmuster. Dieselben Beispielterme erzeugen also sehr viel mehr Affixmuster als graphische Muster. Andererseits passen die Affixmuster sehr viel seltener auf Terme im zu bearbeitenden Text.

Jeder Beispielterm kann ein Muster erzeugen, aber identische Muster werden nur einmal gespeichert. Die Anzahl der Beispielterme, die ein identisches Muster erzeugen, spielt also keine Rolle. Mit anderen Worten werden Regel und Ausnahme gleich behandelt. Wenn die Beispielterme also in einer Weise von der Verteilung der Terme im Text abweichen (z.B. fast ausschließlich Nominalphrasen und nur einige wenige Verbalphrasen), dann werden die extrahierten Terme doch der Verteilung im Text (ein Teil Nominalphrasen und ein Teil Verbalphrasen) und nicht der Verteilung der Beispielterme gleichen.

Allgemein kann man sagen, dass Parameter die Zeichenketten intern oder in ihrem Zusammenhang beschreiben. Die gängigsten internen Parameter beziehen sich darauf, ob die Zeichenketten Satzzeichen, Groß- oder Kleinbuchstaben, Affixe oder Suffixe, Vokale oder Konsonanten, jeweils in bestimmter Anzahl und Position, enthalten. Die Parameter zum Zusammenhang der Zeichenketten betrachten deren relative und absolute Position im Text.

Das Programm zerlegt die Beispiele nach den vorgegebenen Parametern (z.B. Groß- oder Kleinschreibung am Wortanfang) in konkrete Muster (klein-klein-klein). Jedes Beispiel erzeugt durch jeden Parameter ein Muster. Da jedes Muster aber nur einmal gespeichert wird und die Varianz meist nicht besonders groß ist (es gibt nur 8 Muster bei grafischer Varianz von Dreiworttermen), kommen durch neue Beispiele oft keine neuen Muster hinzu und es müsste bald eine Sättigung eintreten. Dies überprüft der Versuch unter 8.2.

### 4.3. Verbesserung der beispielbasierten TE

Durch Lernschleifen können sowohl neue Beispiele als auch neue Aufbaumuster generiert werden. Beispiele und Regeln können mit einem Korpus überprüft werden.[213]

**Zusätzliche Filter** können Effizienz und Qualität der Termextraktion steigern.[214] Man kann etwa die häufigsten Wörter eines Hintergrundkorpus als Funktionswörter definieren und die Regel aufstellen, dass Funktionswörter nie die Randstellung eines Terms einnehmen. Durch einen weiteren Filter kann man von vornherein ausschließen, dass Termkandidaten Satzzeichen (z.B. Punkte oder

[212] Die ladinischen Beispiele sind aus dem Gadertalerischen.

[213] Quasthoff U., Biemann C., Wolff C. (2002), Named Entity Learning and Verification: Expectation Maximization in Large Corpora, S. 8-14 in: Roth D., Bosch A. van den (Hrsgg.) (2002), Proceedings of the 6th Workshop on Computational Language Learning (CoNLL), 2002 Taipei.

[214] Merkel M., Nilsson B., Ahrenberg L. (1994), A phrase-retrieval system based on recurrence, S. 99-108 in: Proceedings of the Second Annual Workshop on Very Large Corpora (WVLC-2), Kyoto 1994, http://www.ida.liu.se/~magme/publications/kyoto--94.pdf : 13.9.2004.

Klammern) enthalten. In unseren Versuchen (s.u.) zeigte auch ein Filter für zu kurze und zu lange Termkandidaten positive Wirkung. Die gewünschte Länge eines Termkandidaten wurde mit ±3 Standardabweichungen von der mittleren Länge der Beispielterme festgelegt.

4.4. *ranking*

Die extrahierten Termkandidaten müssen nach ihrer Termwahrscheinlichkeit geordnet werden. Das *ranking* (englisch *to rank*: reihen, in eine Rangordnung bringen) von Indexterme ist besonders dann ausschlaggebend, wenn man sich auf die ersten Terme als Indexterme beschränkt. Wenn *recall* und *precision* schlecht sind, d.h. wenn sich nur wenige echte Terme unter vielen falschen Termen finden, dann kann man doch noch zu einer guten Ergebnisliste gelangen, wenn es gelingt, die wenigen richtigen Terme auf die vordersten Listenplätze zu schieben und dann nur die besten Termkandidaten verwendet. Man erhält statt allen extrahierten Termkandidaten nur diejenigen als Ergebnis, die am wahrscheinlichsten auch Terme sind. So wird aus einer ‚schlechten' Termextraktion durch geschicktes *ranking* noch eine gute.

Die Ergebnislisten können bei allen Verfahren am Schluss geordnet werden. Beim *Information Retrieval* wird zum Ordnen von Ergebnisdokumenten nach ihrer Relevanz wird das TF.IDF-Maß verwendet, also die Häufigkeit im jeweiligen Text dividiert durch die Häufigkeit in allen Dokumenten. Wenn das Ordnungskriterium dasselbe ist wie bei der Extraktion, dann ergibt sich aber keine neue Ordnung. Statistische Verfahren bekommen durch statistische Maße keine neue Information, da die Termkandidaten bereits nach statistischen Kriterien ausgewählt wurden. Daher bleibt dem statistischen Ansatz nur das Ermitteln eines Schwellenwerts, bei dem die Liste abgeschnitten wird.

Allerdings ist das Ermitteln der Reihenfolge auch bei anderen Ansätzen nicht unproblematisch, denn für statistische Messungen fehlt es oft an einem hinreichend großen Hintergrundkorpus, insbesondere wenn es um entsprechend seltene Mehrwortausdrücke geht. Beim TF.IDF-Maß stünde dann häufig 0 im Nenner, wodurch man keine geeignete Vergleichsgröße hätte. Dann muss die sehr einfache Variante Wortlänge mal Wortfrequenz zum Ordnen verwendet werden.[215]

## 5. Vergleich der drei TE-Ansätze

Die drei Ansätze können außerdem nach ihren Voraussetzungen, Sprachunabhängigkeit, Kontrollierbarkeit des Verfahrens und Wiederverwendbarkeit der vorausgesetzten Ressourcen unterschieden werden.

5.1. Voraussetzungen

Die **Voraussetzung** für statistische Verfahren sind große Textmengen in elektronischer Form und in den Sprachen der Dokumente. In den gebräuchlicheren Sprachen stehen genügend große Textkorpora zur freien Verfügung.

Beim beispielbasierten Ansatz kann mit einem bis zwei Beispielen begonnen werden. Gefundene Lösungen können als weiteres Beispiel hinzugefügt werden (Lernschleife), sollten aber vorher verifiziert werden, weil sich sonst die Fehler fortpflanzen. Im unten beschriebenen Versuch ohne Lernschleife waren 100 Beispiele ausreichend.

Bei regelbasierten Verfahren werden von Experten linguistische Regeln ausgewählt. Dies setzt voraus, dass ein muttersprachlicher Linguist konzeptuelles und sprachliches Wissen formalisiert. Damit ist dieser Ansatz bezüglich des eingesetzten Wissens am anspruchsvollsten, bezüglich der empirischen Daten aber am anspruchslosesten.

[215] Zum *ranking* als Vergleichsmaß siehe 7.3.

5.2. Sprachabhängigkeit

Der statistische Ansatz ist insoweit **sprachabhängig**, als er Korpora in der Dokumentsprache erfordert. Der beispielbasierte Ansatz benötigt Beispiele in der Dokumentsprache sowie formale Ansatzpunkte für die Regelgenerierung, z.B. dass Satzzeichen Sinngrenzen darstellen oder dass die Sprache am Wortende flektiert. Nur der regelbasierte Ansatz ist voll sprachabhängig.

5.3. Kontrollierbarkeit des Verfahrens

Mit **Kontrollierbarkeit des Verfahrens** ist die gezielte Manipulierbarkeit der Ergebnisse durch die Eingabe gemeint. Der statistische Ansatz ist sehr manipulationsresistent, weil einzelne Manipulationen in der Menge der Daten untergehen. Der beispielbasierte Ansatz kann zwar leicht manipuliert werden, etwa durch Eingabe anderer Beispiele (siehe im Versuch unten die Eingabe von Fachbegriffen und allgemeinen Wörterbucheinträgen), aber nicht gezielt, da der Effekt kaum voraussehbar ist. Beim regelbasierten Ansatz kann der Experte einzelne Regeln mit spezifischer Funktion hinzufügen oder verändern und das Ergebnis damit gezielt verbessern. Zum einen können solche Manipulationen aber zu einer Überspezialisierung auf den Trainingsdatensatz führen. Zum anderen bilden die intellektuell gefundenen sprachlichen Regeln aufgrund ihrer Anzahl, komplexen Formalisierung und Interferenz ein Geflecht, das letztlich selbst für Experten weder anschaulich noch kontrollierbar bleibt.

5.4. Wiederverwendbarkeit

Unterschiede bestehen auch hinsichtlich der **Wiederverwendbarkeit** der einzelnen Komponenten, also der Eingabedaten und der verwendeten Programme. Die bei statistischen Verfahren verwendeten Texte stammen meist aus Korpora, die von Korpuslinguisten intensiv für die verschiedensten Untersuchungen genutzt werden. Beispiele werden meist Teil des Projektergebnisses und sind daher keine zusätzliche Arbeit. Linguistische Regeln werden hingegen in der Praxis kaum je wieder verwendet. Programme des statistischen und beispielbasierten Ansatzes können eher wieder verwendet werden als diejenigen des regelbasierten Ansatzes.

5.5. Herkunft der Information

Eine andere Möglichkeit der Einteilung ergibt sich nach der **Herkunft der Information** über die Termkandidaten. Wenn die Information im Termkandidaten selbst steckt (morphologische, syntaktische oder semantische Informationen), dann wird der Ansatz intrinsisch (*intrinsic approach*) genannt, wenn die Information von außerhalb des Termkandidaten gewonnen wird, extrinsisch (*extrinsic approach*). Extrinsische Information kann syntagmatisch (syntaktische oder kontextuelle Information) oder paradigmatisch sein (Information über die Beziehungen zwischen Termkandidaten und Termen).

**Tabelle 9: Übersicht über Termextraktionsmethoden**

| Ansatz | | | **Methode** | **verwendet von:** |
|---|---|---|---|---|
| **Linguistischer Ansatz** | intrinsisch | | POS-*tagging und chunking* | Bourrigault/Jacquemin[216] |
| | | | Stoppwörter | Merkel Mikael[217] |
| | extrinsisch | syntagmatisch | volles Parsing | Arppe[218]; Soininen et al.[219] |
| | | paradigmatisch | Termvariation | Jacquemin[220] |
| **Statistische Ansatz** | intrinsisch | | *mutual information* | Church Hanks[221] |
| | | | *likelihood ratio* | Hong et al.[222] |
| | extrinsisch | syntagmatisch | nc-Wert | Maynard/Ananiadou |
| | | | Entropie | Merkel/Mikael |
| | | paradigmatisch | c-Wert | Nakagawa[223] |
| | | | *weirdness-ratio* | Brekke et al. |

5.6. Verwendung

Beispielbasierte Ansätze finden insbesondere bei der Textklassifikation und beim Parsen **Verwendung**. In der Textklassifikation werden bereits klassifizierte Dokumente verwendet, zum Parsen Parsingbäume oder einzelne geparste Einheiten. In Versuchen mit Chinesisch konnten mit einem einzigen Parsingbaum 43 % aller Abhängigkeitsbeziehungen erkannt werden,[224] was das besonders gute Verhältnis von Aufwand zu Ergebnis unterstreicht.

[216] Bourigault D., Jacquemin C. (1999), Term extraction and term clustering. An integrated platform for computer-aided-terminology, S. 15-22 in: Proceedings of 9th Conference of the European Chapter of the Association for Computational Linguistics (EACL 1999), EACL Bergen 1999, http://citeseer.nj.nec.com/bourigault99term.html : 5.10.2004.

[217] Merkel M., Mikael A. (2000), Knowledge-lite extraction of multi-word units with language filters and entropy thresholds, S. 737-746 in: Proceedings of the 6th International Conference Content-Based Multimedia Information Access (Recherche d'information assistée par ordinateur - RIAO 2000), Vol. 1, Collège de France Paris 2000, http://citeseer.nj.nec.com/merkel00knowledgelite.html : 4.10.2004.

[218] Arppe A. (1995), Term extraction from unrestricted text, short paper in: Koskenniemi K. (Hrsg.) (1995), Proceedings of the 10th Nordic Conference of Computational Linguistics (Nordiska datalingvistdagarna: NoDaLiDa 1995), Helsinki 1995, http://www.lingsoft.fi/doc/nptool/term-extraction.html : 5.10.2004.

[219] Soininen P., Voutilainen A., Tapanainen P. (1999), An experiment in automatic term extraction, S. 234-240 in: Sandrini P. (Hrsg.) (1999), Proceedings of the 5th International Congress on Terminology and Knowledge Engineering (TKE) 1999, TermNet Innsbruck 1999.

[220] Jacquemin C. (1999), Syntagmatic and paradigmatic representation of term variation, S. 341-348 in: Proceedings of the 7th Annual Meeting of the Association for Computational Linguistics (ACL) 1999, ACL Maryland 1999, http://citeseer.nj.nec.com/jacquemin99syntagmatic.html : 5.10.2004.

[221] Church W., Hanks P. (1989), Word association norms, mutual information and lexicography, S. 76-83 in: Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics, ACL Vancouver 1989, http://citeseer.nj.nec.com/church89word.html : 5.10.2004.

[222] Hong M., Fissaha S., Haller J. (2001), Hybrid filtering for extraction of term candidates from German technical texts, Poster in: Proceedings of Terminologie et Intelligence Artificielle (TIA), INIST Nancy 2001, http://www.iai.uni-sb.de/docs/term_extract.pdf : 5.10.2004.

[223] Nakagawa H. (2001), Experimental evaluation of ranking and selection methods in term extraction, S. 303–325 in: Bourigault D. et al. (2001), a.a.O. http://www.r.dl.itc.u-tokyo.ac.jp/~nakagawa/academic-res/termrec.pdf : 5.10.2004.

[224] Streiter O., De Luca E. W. (2003), Example-based NLP for Minority Languages: Tasks, Resources and Tools, S. 233-242 in: Proceedings of the Conférence sur le Traitement Automatique des Langues Naturelles (TALN), Association pour le Traitement Automatique des Langues (ATALA) Batz-sur-Mer 2003.

Beispielbasierte Ansätze eignen sich insbesondere für Minderheitssprachen und ähnlich ungünstige Situationen, in denen weder Korpora noch muttersprachliche Computerlinguisten eingesetzt werden können.[225]

## 6. TE zur Indexierung

### 6.1. Indexieren als Teilaufgabe des *Information Retrieval*

Allgemeines Ziel des *Information Retrieval* (IR) ist es, ausschließlich alle für den Suchenden relevanten Dokumente auf vage Suchanfragen anzugeben. Dazu müssen die Dokumente und das Informationsbedürfnis in einer Weise repräsentiert werden, dass die Ähnlichkeit der Repräsentationen verglichen werden kann.[226]

Voraussetzung für die Ähnlichkeitsmessung ist ein gemeinsames Maß von Dokumenten und Informationsbedürfnis. Dokumente sind Zeichenketten, das Informationsbedürfnis menschlicher Benutzer ist wissensbezogen. Die Überbrückung der Kluft zwischen Zeichen und Bezeichnetem ist ein sehr komplexes Problem, mit dem sich ganze Wissenschaftszweige beschäftigen.[227] Im IR gibt es Ansätze, die dieses Problem ganz ausklammern bis hin zu solchen, die dieses Problem komplett zu lösen versuchen.

Dem entsprechend kann man auch beim IR zwischen semasiologischen und onomasiologischen Ansätzen unterscheiden und die Repräsentationen von Dokumenten oder Suchanfragen sind eher zeichen- oder wissensorientiert. Das zeichenorientierte Extrem ist eine Volltextsuche, die das Retrievalproblem wegen der Variabilität von Sprachen oft nicht zufrieden stellend löst. Das wissensorientierte Extrem ist eine Ontologie[228], bei der die Bezeichnung einer Wissensklasse im Dokument selbst kaum auftaucht. Keines dieser IR-Systeme bearbeitet die in den Dokumenten enthaltenen Zeichen, denn eine Volltextsuche verwendet alle Zeichen, eine Ontologie gar keine.

Die Grafik 32 soll einen Gesamtüberblick über den Zusammenhang der in diesem Kapitel angesprochenen Methoden, Verfahren und Hintergründe geben und nimmt daher auf einiges Bezug, das im Text an späterer Stelle ausführlich besprochen wird. Die linke Seite ist die Seite der Zeichen, wo von einem Text ausgehend mit statistischen Verfahren und der Eliminierungsstrategie gearbeitet wird. Diese Seite ist sprachabhängig, bezieht wenig sprachliches und konzeptuelles Wissen mit ein, sie arbeitet wenig mit Formalisierungen und Abstraktionen und die Textrepräsentationen sind weniger leicht intuitiv verstehbar. Kontinuierlich zur rechten Seite gehend verhält es sich dort umgekehrt.

---

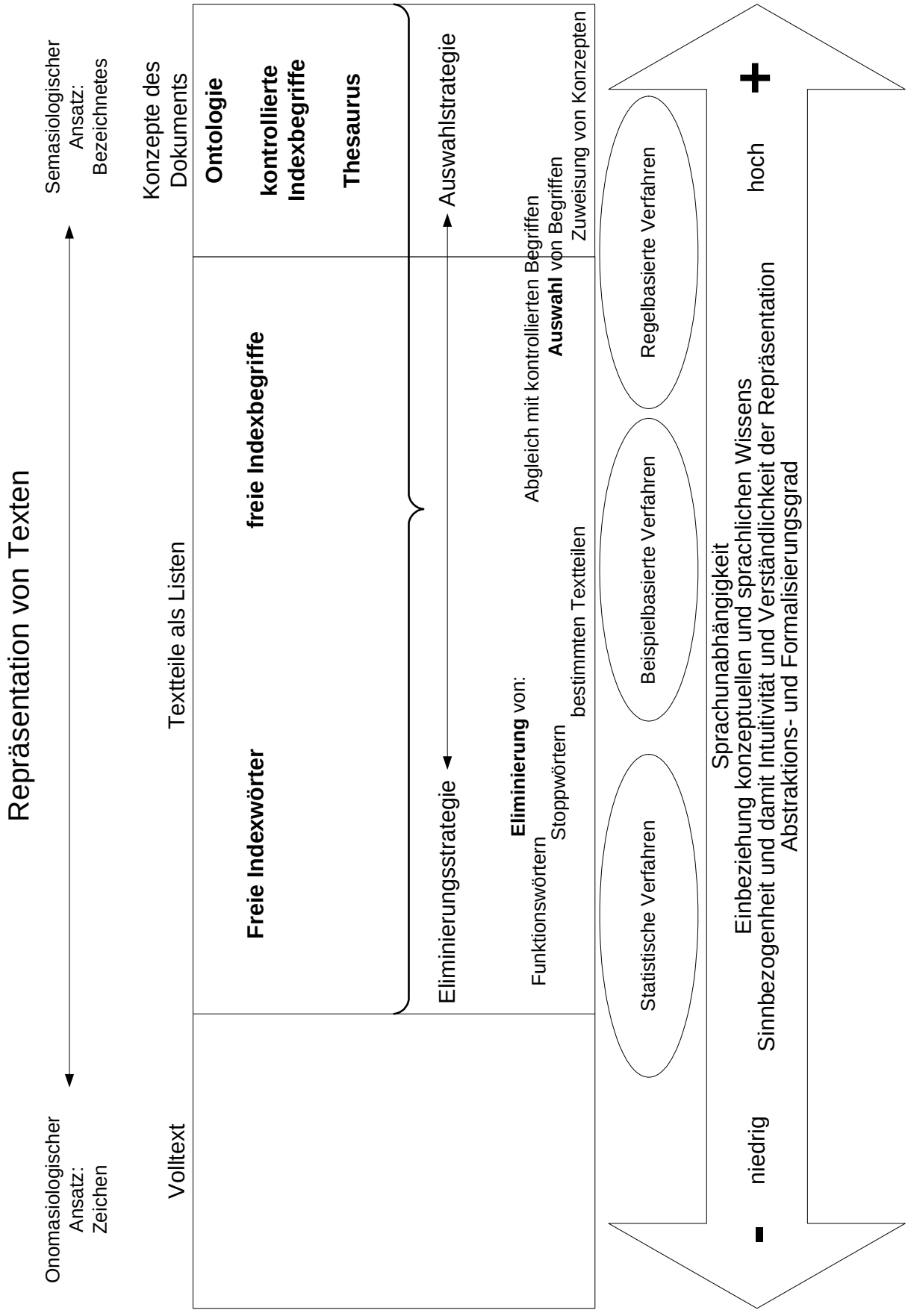
[225] Streiter O., Zielinski D., Ties I., Voltmer L. (2002), Example-based Term Extraction for Minority Languages: A case-study on Ladin, in: Tagungsband von Soziolinguistica Y Language Planning, SPELL St. Ulrich (Italien) noch in Vorbereitung, http://dev.eurac.edu:8080/autoren/publs/termex5.pdf : 5.10.2004.

[226] Van Rijsbergen C. J. (1983), Information Retrieval, Butterworths London 1983, Kapitel 1.

[227] Die Semantik, ein Teilbereich der Semiotik, befasst sich mit der Relation zwischen Zeichen und bezeichnetem Objekt, während die Übersetzungswissenschaft mit der praktischen Seite des Problems konfrontiert ist. Im Alltag wird das Problem etwa bei Suchanfragen in mehrsprachigen Umgebungen deutlich, wenn dieselben Zeichen unterschiedliche Bedeutung in mehreren Sprachen haben.

[228] Ontologie bedeutet hier die explizite formale Spezifikation einer gemeinsamen Konzeptualisierung. Gruber T. R. (1993), Towards Principles for the Design of Ontologies Used for Knowledge Sharing, in Guarino N., Poli R. (Hrsgg.) (1993), Formal Ontology in Conceptual Analysis and Knowledge Representation, Kluwer Academic Publishers Deventer (NL) 1993.

**Grafik 32: Das Form-Inhalt-Kontinuum der Wissensrepräsentation**



Viele Textdokumentsammlungen sind neben einer Volltextsuche über ein weiteres, hybrides, zeichenverarbeitendes IR-System zugänglich. Zeichenorientierte Ansätze verwenden eher eine Eliminierungsstrategie, wissensorientierte Ansätze eher eine Auswahlstrategie.

Bei der **Eliminierungsstrategie** werden von allen logisch möglichen Zeichenketten eines Dokuments ausgehend diejenigen Zeichenketten eliminiert, die zum Repräsentationszweck nichts beitragen, etwa weil ihre Unterscheidungskraft gering ist. Ein Beispiel für eine solche Eliminierung sind Stoppwortlisten. In aller Regel bleiben dabei relativ viele Zeichenketten übrig.

Bei der **Auswahlstrategie** werden vorher feststehende Bezeichnungen im Text gesucht oder dem Text zugeordnet. Meist beschränken sich solche Repräsentationen auf wenige Schlagwort-, Thesaurus- oder Indexbegriffe[229].

In der Tendenz ist die Eliminierungsstrategie eher *recall*-orientiert und kommt mit weniger explizitem sprachlichen Wissen aus, während die Auswahlstrategie eher *precision*-orientiert ist und sich eher auf linguistische Verfahren stützt (etwa zur Erstellung eines kontrollierten Vokabulars).

### 6.2. Textverarbeitende Indexierungsverfahren

Textverarbeitende Repräsentationsverfahren können nach der Struktur der Eingabe- und Ausgabedaten in drei Gruppen eingeteilt werden: Bei der Eingabe von Regeln zur Ausgabe von Indextermen ist der Input komplexer als der Output (**regelbasierte Verfahren**). Bei der Eingabe von Beispielbegriffen zur Ausgabe von Indextermen ist der Input von der gleichen Komplexität wie der Output (**beispielbasierte Verfahren**). Bei der Eingabe von Text zur Ausgabe von Indextermen ist der Input von geringerer Komplexität als der Output (**statistische Verfahren**).[230]

Diese Dreiteilung gibt an, wie komplex die Vorarbeiten zur Abstraktion und Formalisierung sind. Den höchsten Grad der Verarbeitung und Abstraktion erfordern Regeln und Ausnahmen zur Bildung der gesuchten Einheiten. Solche Regeln müssen in programmlesbarer Weise formalisiert sein. Beispieleinheiten erfordern nur relativ einfache Vorarbeiten und die Formalisierung wird vom Programm nach vorgegebenen Parametern durchgeführt. Einfacher Text braucht gar nicht bearbeitet zu werden und enthält noch die volle Komplexität der Sprache.[231]

Nur der statistische und der beispielbasierte Ansatz kommen ohne explizites linguistisches Wissen aus. Der statistische Ansatz benutzt große Textmengen, um die wichtigen Begriffe herauszufinden. Dem beispielbasierten Ansatz dient die Struktur der Beispiele als Ausgangspunkt für die Suche nach Indextermen.

Mit der **Verständlichkeit** textverarbeitender Indexierungsverfahren ist gemeint, wie intuitiv verständlich die Verbindung zwischen dem Index und den Dokumenten ist. Dies hängt davon ab, wie stark sinnbezogen der Index ist. Regelbasierte Verfahren sollten besser als beispielbasierte sein, die besser als statistische sein sollten. Statistisch kann ein Eigenname ein besonders geeignetes Indexwort sein, an dem aber der Inhalt der damit indexierten Dokumente nicht abzulesen ist.

---

[229] Ein Thesaurus enthält nach ISO 2788, 1986:2 ein kontrolliertes Indexvokabular, dem Konzepte mit eindeutigen Beziehungen zugrunde liegen.

[230] Wissenschaftstheoretisch werden zunächst induktiv Regeln erstellt, die dann deduktiv angewandt werden. Der induktive Schritt wird beim regelbasierten Ansatz vor dem Beginn der Datenverarbeitung durch den Linguisten vollzogen, beim statistischen und beispielbasierten Ansatz ist die Induktion Teil der Datenverarbeitung.

[231] Aus dem Blickwinkel des IR soll das Kernproblem, die Kluft zwischen Zeichen und Bezeichnetem, mit der Eingabe von linguistischen Regeln gelöst werden, während es bei der Eingabe von Text auf die Textverarbeitung verschoben wird. Durch die Eingabe von Beispielen wird die Kluft nicht überbrückt, sondern es werden die Orte anderer Brücken angegeben, um das Auffinden geeigneter Überbrückungsmöglichkeiten zu erleichtern. Beispiele sind in funktioneller Sicht Text, der sich besonders gut zur Findung von Termbildungsregeln eignet.

In der wissenschaftlichen Literatur werden vorherrschend statistische und regelbasierte Ansätze zur automatischen Indexierung durch Termextraktion verwendet. Jacquemin & Bourigault[232] geben den neuesten Stand wieder und unterteilen zum einen danach, ob die Suche nach Termen erstmals stattfindet oder ob eine bestehende Liste erweitert wird, und zum anderen danach, ob es einen Numerus Clausus für die Indexterme gibt oder ob grundsätzlich jeder Term in Frage kommt.

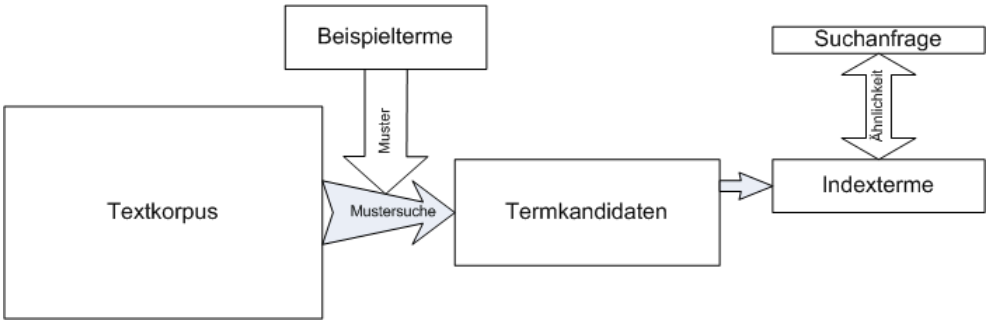
**Tabelle 10: Einteilung von Indexierungsmethoden**

| | **Terme vorhanden** | **Keine Terme vorhanden** |
|---|---|---|
| **Erwerb von Termkandidaten** | erweiternd | originär |
| **Erwerb von Indexermen** | Zuordnung kontrollierter Indexterme | Erstellung eines freien Indexes |

In diesem Schema ist die beispielbasierte Indexierung nicht eindeutig einzuordnen. Zwar müssen Beispielterme vorhanden sein, allerdings genügen sehr wenige Beispiele, die manuell erstellt oder aus anderen Projekten (z.B. Internetglossaren) übernommen werden können. Zur manuellen Erstellung von Beispielen ist ganz elementares Wissen über die Sprache ausreichend. Die Beispiele werden außerdem zu Beispielmodellen weiter verarbeitet und kommen nicht in die Liste der Termkandidaten. Der Erwerb von Termkandidaten ist bei der beispielbasierten Termextraktion daher originär.

In der Regel werden die extrahierten Termkandidaten gleich als Indexterme in einen freien Index übernommen. Möglich wäre ein Abgleich mit feststehenden Termen zu einem kontrollierten Index. Dazu wäre die Termextraktion aber nicht nötig, weil die Indexterme sofort im Dokumenttext gesucht und indiziert werden können.

**Grafik 33: TE durch Beispielterme und anschließende Indexierung**




Die Grafik 33 legt die Vermutung nahe, dass die Qualität von Termkandidaten für die Indexierung steigen könnte, wenn man Suchanfragen als Beispielterme verwendet.

### 7. Vergleichsmaße für TE und Indexierung

Der konkrete Nutzen einer Termextraktion hängt stark davon ab, wie die Termextraktion in den terminografischen Prozess eingebunden wird. Objektiv vergleichbar werden verschiedene Methoden durch Bewertung der Treffergenauigkeit bzw. Präzision (*precision*), die Vollständigkeit (*recall*), eine Kombination dieser beiden zu einem Mittelwert (*mean* oder *F-measure*) sowie die Bewertung der Rangliste der Termkandidaten (*ranked recall*).

[232] Jacquemin C., Bourigault D., (2003), Term Extraction and Automatic Indexing, in: Handbook of Computational Linguistics, Mitkov R. (Hrsg.) (2003), Oxford University Press Oxford 2003, beschreiben die verschiedenen Ansätze und einsprachigen Termextraktionsanwendungen TERMINO, LEXTER, ACABIT, Xtract, ANA sowie die Termerkennungs- und Indexierungsanwendungen FASIT, CLARIT, TTP, COB, COPSY and FASTR.

$$recall = \frac{\#\{\{TC\} \cap \{T_{doc}\}\}}{\#\{T_{doc}\}}$$

Diese Bewertungsmethoden kommen aus dem *information retrieval* und funktionieren formal gesehen auch für die Termextraktion. Bei der Übertragung der Bewertungsmethoden muss jedoch stets im Auge behalten werden, dass das Ziel einer Informationssuche ein anderes ist als beim Finden noch nicht beschriebener Terminologie (*term discovery*). Die Schwerpunkte sind daher anders gesetzt und die Ergebnisse müssen anders interpretiert werden.

### 7.1. *recall*

Die **Vollständigkeit** (*recall*) einer Termextraktion ist das Verhältnis von extrahierten Termkandidaten (TC = *term candidates*) zu vorhandenen Termen ($T_{doc}$). Bei einem *recall* von 80 % bleiben 20 % der Termkandidaten unentdeckt. Für eine Informationssuche sind nicht entdeckte Informationen fatal, weil diese Informationen dem Nutzer verloren gehen. Bei einer Termerkennung kann es hingegen akzeptabel sein, auf einen Teil der möglichen Terme zu verzichten, weil die nicht entdeckten Terme später noch gefunden werden können. Wenn die gefundenen Terme der Datenbank hinzugefügt wurden, kann erneut eine Termextraktion über denselben Text laufen und je nach Einbindung dieser neuen Informationen können weitere Terme gefunden werden (*bootstrapping*). Noch einfacher wäre es aber, diese Terme, die als Fachwörter per Definition auch in anderen Texten vorkommen, durch Termextraktion in weiteren Texten aufzuspüren.

Oft wird der *recall* gar nicht berechnet, weil dazu alle Terme eines Textes vollständig bestimmt werden müssten, was arbeitsintensiv und im Einzelfall schwierig ist. Man arbeitet daher oft mit relativer Vollständigkeit.

### 7.2. *precision*

Die **Treffergenauigkeit** (*precision*) ist das Verhältnis der erkannten Termini zu den extrahierten Termkandidaten.

$$precision = \frac{\#\{\{TC\} \cap \{T_{doc}\}\}}{\#\{TC\}}$$

Bei einer *precision* von 80 % sind vier von fünf Termkandidaten Terme. Je geringer die Treffergenauigkeit ist, umso wichtiger ist die manuelle Nachbearbeitung der Ergebnisse. Bei der Informationssuche wird die Bedeutung von Dokumenten (Information) gesucht, während bei der Terminologieerkennung eine bestimmte Formalisierung (Term) von Information (Bedeutung eines Terms) gesucht wird. Je stärker die Formalisierung ist, umso höher wird die Treffergenauigkeit sein. Wenn nur Nominalphrasen als Terme zugelassen werden, dann hat die Terminologieerkennung einen wesentlichen Vorteil gegenüber der Informationssuche. Selbstverständlich muss auch beim Vergleich verschiedener Termextraktionssysteme die jeweilige Formalisierung von Begriffen berücksichtigt werden.

Vollständigkeit und Treffergenauigkeit sind indirekt voneinander abhängig. Werden alle möglichen Wortkombinationen extrahiert, dann ist die Vollständigkeit 1 und die Treffergenauigkeit nimmt den niedrigsten Wert an (nicht Null, sondern abhängig von der Anzahl der Terme). Die Treffergenauigkeit maximiert man dadurch, dass nur der Termkandidat mit dem besten Wert extrahiert wird. Damit wird gleichzeitig die Vollständigkeit minimiert, denn ist dieser Termkandidat kein Term, dann ist die Treffergenauigkeit Null. Daher werden Vollständigkeit und Treffergenauigkeit zur Evaluierung oft zu einem gemeinsamen Wert (F-Wert, *F-score* oder *mean*) kombiniert:

$$F\text{-}score = \frac{2 \cdot precision \cdot recall}{precision + recall} = \frac{1}{\alpha \frac{1}{precision} + (1-\alpha)\frac{1}{recall}} = \frac{2 \cdot \#\{\{TC\} \cap \{T_{doc}\}\}}{\#\{T_{doc}\} + \#\{TC\}}$$

mit α = Gewichtung zwischen Vollständigkeit und Treffergenauigkeit.

Da der Zusammenhang nicht linear ist, gibt es einen optimalen F-Wert. Auch wenn die Veränderung des F-Werts oft nur gering erscheint, sollte beim Vergleich verschiedener Termextraktionsmethoden darauf geachtet werden, dass der jeweils beste F-Wert verglichen wird, auch wenn der Schwerpunkt einer Termerkennung auf der Treffergenauigkeit liegt.

### 7.3. *ranked recall*

Die Treffergenauigkeit wird für alle Termkandidaten berechnet. Bei statistischen Ansätzen werden aber Ergebnisse für alle möglichen Kombinationen berechnet. Ergebnisse unterhalb eines Schwellenwertes werden dann außer Acht gelassen bzw. die Trefferliste wird unten abgeschnitten. Damit wird die Treffergenauigkeit zur Kunst des Abschneidens. Deshalb ist teilweise dazu übergegangen worden, bei der Termextraktion die Rangfolge zu bewerten. Eine Termextraktion, die acht Terme findet, aber an den Stellen 3 bis 10, ist schlechter als eine Termextraktion, die die Terme an die Stellen 1 bis 8 auflistet. Dies ist die Fähigkeit, die guten von den schlechten Termkandidaten zu trennen, der *ranked recall*. Wenn $r_i$ der Listenplatz des *i*-ten Termkandidaten mit $TC \middle| TC \in \{\{TC\} \cap \{T_{doc}\}\}$ ist, dann ist der

*ranked recall* definiert als: $ranked\ recall = \frac{\sum_i^n i}{\sum_i^n r_i}$.

In einer Liste von zehn Termkandidaten, in der die acht Terme auf den Plätzen drei bis zehn sind,

ist der *ranked recall*: $\frac{1+2+3+4+5+6+7+8}{3+4+5+6+7+8+9+10} \approx 0.69$.

Sind die acht Terme auf den ersten acht Plätzen,

dann ist der *ranked recall*: $\frac{1+2+3+4+5+6+7+8}{1+2+3+4+5+6+7+8} = 1$.

Hier ist also jede einzelne Platzierung ausschlaggebend, und zwar umso stärker, je weiter oben der Fehler gemacht wird, weil dadurch alle späteren Termkandidaten einen Platz absacken und einen höheren Platz einnehmen.

## 8. Experimente zur Termextraktion mit Beispieltermen

Mit diesem theoretischen Aufbau wurde die beispielbasierte Termextraktion bei einem ladinischen[233] Rechtstext ausprobiert. Es handelte sich um eine Gemeindeordnung aus Gröden, die 994 Wörter enthielt und in der eine Terminografin mit der traditionellen Methode 113 Terme fand. Die Aufgabe der automatischen Termextraktion war es nun, dieses Ergebnis automatisch zu erzeugen, also dieselben 113 Terme herauszufinden.

[233] Ladinisch ist eine rätoromanische Sprache. Die Variante Grödnerisch wird im Grödnertal in den italienischen Dolomiten von ca. 5000 Personen gesprochen.

8.1. Beispielbasierte Auswahl gegen statistische Eliminierung

Für die Eliminierungsstrategie wurden zunächst alle 19019 möglichen Wortkombinationen erzeugt, die natürlich einen *recall* von 1 haben (Tabelle 4 erste Zeile). Schritt für Schritt wurde nun versucht, mit einfachen Regeln die Wortkombinationen zu reduzieren ohne eine der 113 richtigen zu verlieren. Die Regeln waren: Keine Satzzeichen im Termkandidaten (Tabelle 4 Zeile 2), keine Funktionswörter[234] als erstes oder letztes Wort (Zeile 3) und keine extreme Längenabweichung (Zeile 4).

Um an eine Liste von Funktionswörtern zu gelangen, wurde ein kleines ‚Hintergrundkorpus' mit zwanzig Texten erstellt und die häufigsten Wörter herausgesucht. Dahinter steht die Annahme, dass die häufigsten Wörter Funktionswörter sein werden.[235] Die dritte Regel legt fest, dass die Länge der Termkandidaten nicht extrem von den Beispielen abweichen darf. Dafür wurde die Längenvarianz auf maximal drei Standardabweichungen begrenzt.[236]

Nach Anwendung dieser Regeln ist der *recall* mit fast 94 % weiterhin gut, obwohl doch einige Terme verloren gingen, die Funktionswörter in Randstellung enthielten.

**Tabelle 11: Termextraktion mit einfachen Methoden**

| Methode | #{TC} | *recall* | *precision* | *mean* | *ranked recall* |
|---|---|---|---|---|---|
| **ohne** | 19019 | 1 | 0,0056 | 0,011 | 0,011 |
| **kein Satzzeichen** | 8023 | 1 | 0,0134 | 0,026 | 0,0179 |
| **Funktionswörter** | 6289 | 0,946 | 0,016 | 0,033 | 0,030 |
| **Längenabweichung** | 2419 | 0,9375 | 0,044 | 0,084 | 0,055 |
| **addierte Muster** | 489 | 0,848 | 0,202 | 0,326 | 0,388 |

Für die Auswahlstrategie wurden zwei Gruppen von Mustern erzeugt, und zwar Groß- und Kleinschreibungsmuster und grafische Muster. Jedes Wort und jede Kombination von Wörtern aus der Gemeindeordnung, die mit einem Muster aus jeder der zwei Gruppen übereinstimmte, ist Termkandidat. So wurden 489 Termkandidaten erzeugt (letzte Zeile der Tabelle 4).

Auch die Auswahlstrategie erreicht einen recht guten *recall* von fast 85 %, braucht dazu aber nur knapp 500 Termkandidaten[237], was sich ganz deutlich im *mean* und *ranked recall* niederschlägt.

Der beispielbasierte Ansatz findet vier von fünf Termen, aber nur jeder vierte gefundene Termkandidat ist ein Term. Mit anderen Worten wird das *unithood*-Problem hervorragend gelöst, während die *termhood*-Aufgabe noch verbessert werden muss.

8.2. Ordnung der Termkandidaten

Die Termkandidaten wurden nach ihrer Wortlänge mal der Wortfrequenz im Dokument geordnet. Trotz dieser geradezu primitiven Erstellung einer Reihenfolge ergab sich eine bemerkenswert gute Rangordnung, denn bei diesem und auch allen anderen Versuchen waren die ersten fünf Termkandidaten stets Terme und unter den ersten zehn Termkandidaten höchstens ein Nichtterm. In diesem Fall handelte es sich um die Abkürzung für Artikel „art", also einen rechtlichen Fachbegriff,

[234] Funktionswörter (auch Synsemantikum oder Strukturwort) sind Wörter ohne eigene lexikalische Bedeutung, die vor allem syntaktisch-strukturelle Funktion erfüllen, wie z.B. Artikel, Präpositionen und Konjunktionen. Glück H. (2000), Metzler Lexikon Sprache, Stichwort 'Funktionswort' S. 226, Metzler Stuttgart 2000.

[235] Merkel M., Nilsson B., Ahrenberg L. (1994), A phrase-retrieval system based on recurrence, S. 43-56 in: Proceedings of the Second Annual Workshop on Very Large Corpora (WVLC-2) Kyoto 1994.

[236] Die Standardabweichung ist das bei weitem meistgebrauchte Streuungsmaß und errechnet sich aus der Quadratwurzel der Varianz in den Daten. Die Varianz ist das Quadrat der mittleren Abweichung vom arithmetischen Mittel.

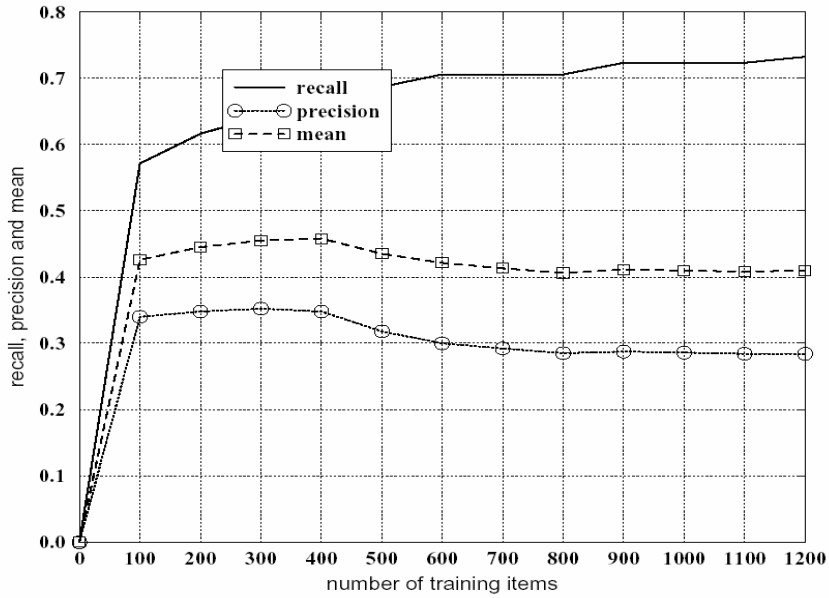
[237] Die 489 Termkandidaten ergeben sich aus der Addition aller Termkandidaten, die aus Fachbegriffen und aus einem Wörterbuch erzeugt wurden (vergl. Tabelle 5: 299+322) unter Ausschluss aller 132 doppelten (621-132=489). Die Kombination vor der Erzeugung von Mustern ergab hingegen nur 390 TC (vergl. letzte Zeile in Tabelle 5).

der für eine Satzung durchaus charakteristisch ist. Die Benennung ist auch in der Datenbank BISTRO enthalten, so dass man diesem Termkandidaten im Nachhinein Termqualität zugestehen kann. Dieses Beispiel zeigt aber deutlich, wie schwierig die Bewertung einer TE ist, weil sie von der jeweiligen Definition eines Terms abhängt. Die ‚Termhaftigkeit' (termhood) scheint ähnlich wie die Fachlichkeit (vergl. Kapitel 1) unscharfe Grenzen zu haben. Jedenfalls kann angenommen werden, dass die Zeichenkette „art" von einem Informationssuchenden benutzt wird, z.B. als „Art. 1 Gemeindeordnung". Dieser Termkandidat würde sich somit zumindest als Indexterm eignen.

8.3. Versuche zur Sättigung mit Beispieltermen

Untersucht wurde *recall*, *precision* und *mean*-Wert auch in **Abhängigkeit von der Anzahl der Eingabebegriffe**. Grafik 34 zeigt, dass mit 100 Beispielen bereits die höchste Qualität der Termextraktion erreicht wird, weil bei zunehmenden Beispielen der höhere *recall* mit niedrigerer *precision* einhergeht. Die Erklärung dafür könnte sein, dass die zuerst generierten Regeln mit statistisch höherer Wahrscheinlichkeit die häufig vorkommenden Terme beschreiben, während die später noch hinzukommenden Regeln nur noch die selteneren Termbildungen beschreiben und dabei relativ häufiger auch auf Nichtterme passen. Das „Hintergrundrauschen" in Form von zufällig auf die Termmodelle passenden Mustern fällt bei selten passenden Termmodellen stärker ins Gewicht. Der *recall* steigt zwar weiter an, die Qualität des Gesamtsystems (*mean*) sinkt aber aufgrund der überproportional hereinkommenden falschen Termkandidaten.

**Grafik 34: *recall, precision* und *mean*-Wert in Abhängigkeit von der Anzahl der Beispiele**




8.4. Fachsprachliche gegen allgemeinsprachliche Beispiele

Als Beispielterme wurden drei Mengen ausprobiert. Die ersten 1225 Beispiele waren traditionell ausgewählte Terme aus der Termdatenbank, die zweite Menge waren bearbeitete Wörterbucheinträge und die dritte Menge ein Mischung aus beiden. Mit oben genannten Formalisierungsmöglichkeiten wurden Muster erzeugt und die Ähnlichkeit wurde als gegeben erachtet, wenn ein Termkandidat mindestens mit einem graphischen Muster (Groß- und Kleinschreibung) und einem Affixmus-

ter übereinstimmte.[238] Die drei Mengen variieren, wie Tabelle 12 zeigt, stark in der Größe, in *recall* und *precision*, aber die *mean*-Werte liegen doch erstaunlich nah beieinander.

**Tabelle 12: TE mit unterschiedlichen Termbeispielen**

| Methode | Anzahl extrahierter Termkandidaten | *recall* | *precision* | *mean* |
|---|---|---|---|---|
| Fachbegriffe | 299 | 0.732 | 0.284 | 0.410 |
| Allgemeinbegriffe | 322 | 0.750 | 0.269 | 0.396 |
| Kombination | 390 | 0.839 | 0.248 | 0.386 |

Im Folgenden wurde mit allen 489 Beispielen weitergearbeitet, die durch Muster erzeugt werden konnten, weil so noch am wenigsten richtige Termkandidaten verloren waren (vergl. Tabelle 4 letzte Zeile: *recall* von 84,8 %). Um die Präzision zu verbessern, wurden die Termkandidaten entfernt, die Satzzeichen enthielten, mit einem Funktionswort begannen oder endeten, oder stark in der Länge abwichen. Wie Tabelle 13 zeigt, konnte vor allem die Funktionswortregel die Anzahl der Termkandidaten reduzieren und den *mean*-Wert erhöhen.

**Tabelle 13: Termextraktion mit kombinierten einfachen Methoden**

| Methode | Anzahl | *recall* | *precision* | *mean* |
|---|---|---|---|---|
| Muster | 489 | 0,848 | 0,202 | 0,326 |
| Muster + kein Satzzeichen | 489 | 0,848 | 0,202 | 0,326 |
| Muster + Funktionswörter | 390 | 0,839 | 0,204 | 0,386 |
| Muster + Längenabweichung | 477 | 0,839 | 0,203 | 0,328 |

Nun wurde der Versuch mit der *weirdness-ratio* (w-r) für Einwortterme durchgeführt (Tabelle 14). Dabei gehen von vornherein 46 % aller Terme verloren, weil Rätoromanisch eine analytische Sprache ist. Derselbe Ansatz kann aber für synthetische Sprachen gut funktionieren. Alle Methoden finden gleich viele richtige Terme (der *recall* in Tabelle 14 ist stets 54,4 %), daher ist die restriktivste Methode hier die beste.

**Tabelle 14: Termextraktion mit der *weirdness-ratio* (w-r)**

| Methode | Anzahl | *recall* | *precision* | *mean* | *ranked recall* |
|---|---|---|---|---|---|
| w-r | 345 | 0,544 | 0,188 | 0,280 | 0,363 |
| w-r + Muster | 312 | 0,544 | 0,210 | 0,303 | 0,415 |
| w-r + Längenabweichung | 316 | 0,544 | 0,205 | 0,298 | 0,400 |
| w-r + Muster + Längenabweichung | 302 | 0,544 | 0,215 | 0,308 | 0,416 |
| w-r + Funktionswörter | 281 | 0,544 | 0,225 | 0,318 | 0,367 |
| w-r + Funktionswörter + Muster | 250 | 0,544 | 0,254 | 0,346 | 0,404 |
| w-r + Funktionswörter + Muster + Längenabweichung | 249 | 0,544 | 0,255 | 0,347 | 0,404 |

Nun wurde der Versuch auch noch für die *mutual information* und Zweiwortterme durchgeführt (Tabelle 15).

[238] Die Termextraktion wurde auf zwölf Sprachen und weitere Sprachkombinationen erweitert und läuft online sowohl über Webseiten als auch selbst geschriebenen Text kostenlos unter der Adresse http://dev.eurac.edu:8080/cgi-bin/index/TermExtract: 5.10.2004.

**Tabelle 15: Termextraktion mit der *mutual information* (MI)**

| Methode | Anzahl | *recall* | *precision* | *mean* | *ranked recall* |
|---|---|---|---|---|---|
| MI | 807 | 0,098 | 0,013 | 0,024 | 0,007 |
| MI + Muster | 160 | 0,098 | 0,063 | 0,074 | 0,064 |
| MI + Muster + Längenabweichung | 69 | 0,098 | 0,144 | 0,110 | 0,144 |

Auch hier war die größte Einschränkung die erfolgreichste Methode. Man kann die Einwortmethode und die Zweiwortmethode als sich ergänzend ansehen. Sie laufen nicht beide über alle 19019 Möglichkeiten, sondern jede nur über einen eigenen, sich nicht überlappenden Teil. Die Ergebnisse einer Kombination von *weirdness-ratio* und *mutual information* ergeben sich daher rechnerisch.

Zusammen konnten 249 + 69 = 318 Termkandidaten extrahiert werden, von denen 63 + 10 = 73 auch Terme waren, so dass sich ein *recall* von 0,646, eine *precision* von 0,230 und ein *mean* von 0,339 errechnet. Für den *ranked recall* müsste man ein Maß einführen, nach dem die Termwahrscheinlichkeit von Einwort- und Zweiworttermen zueinander berechnet wird. Die Kombination von *weirdness-ratio* und *mutual information* schneidet im *recall* und daher auch im *mean*-Wert schlechter ab als die besten Methoden der beispielbasierten Termextraktion.

## 9. TE für Minderheitensprachen in der Literatur

Termextraktion dient meist dem Aufbau von Terminologie. Was für gängige Sprachen ein Geschäft ist, stellt für Minderheitensprachen ein wichtiges Element zur Bildung von Sprachbewusstsein, zur Sprachplanung und -erhaltung dar. Der hohe Aufwand zur Verbesserung von Termextraktionsmethoden kommt selten Minderheitensprachen zugute, denn die meisten Ansätze sind sprachspezifisch und lassen sich nur sehr bedingt oder gar nicht auf Minderheitensprachen übertragen. Das liegt zum einen an den verschiedenen Sprachstrukturen. Währen für germanische und slawische Sprachen Komposita analysiert werden müssen, ist für romanische Sprachen eine analytische Zerlegung nötig. [239]

Die Übertragung von Termextraktionsprogrammen scheitert oft auch an den fehlenden Ressourcen von Minderheitensprachen. Für statistische Ansätze fehlt es an Hintergrundkorpora, für linguistische Ansätze an expliziten morphologischen, syntaktischen oder semantischen Informationen. Darüber hinaus mangelt es fast immer auch an den Linguisten und Computerfachleuten, die solche Ressourcen herstellen könnten. Sprachverarbeitungsressourcen für Minderheitensprachen sind wirtschaftlich nicht interessant und die politischen und wirtschaftlichen Mittel sind daher spärlich gesät.

Zwei Experimente zur Termextraktion in Sprachen mit wenig Sprachverarbeitungsressourcen[240] benutzten die *weirdness-ratio*. Brekke et al.[241] verwenden einen fachsprachlichen Text mit 10.000 norwegischen Wörtern und ein 100.000 Wort großes allgemeinsprachliches Hintergrundkorpus. Die *weirdness-ratio* funktioniert zwar nur für Einwortterme, das mag für Norwegisch mit vielen Kom-

[239] Vergleiche zur Termextraktion für Minderheitensprachen auch Streiter O., Zielinski D., Ties I., Voltmer L. (2003), Term Extraction for Ladin: An Example-based Approach, S. 275-284 in: Tagungsband der Konferenz zum Traitement Automatique des Langues Naturels (TALN) 2003, Batz-sur-Mer 2003.

[240] Norwegisch im Beispiel Fußn. 241 gehört nicht zu den Minderheitensprachen, sondern zu den Sprachen mit wenig Sprachverarbeitungsressourcen. Eine weitere Kategorie vor allem in EU-Förderprogrammen sind die weniger verbreiteten und unterrichteten Sprachen (*less widely used less taught languages*) LWULT, zu denen alle Sprachen außer Englisch, Französisch, Spanisch und Deutsch gehören.

[241] Brekke M. (1996), a.a.O. Fußn. 204.

posita aber passend sein. Auch auf Walisisch wurde die *weirdness-ratio* angewandt, wobei der Fachtext und das Hintergrundkorpus beide je 100.000 Wörter umfassten.[242]

Daille et al.[243] berichten von zwei Experimenten mit Madagassisch.[244] Im ersten Experiment wurde eine statistische, sprachunabhängige Termextraktionsmethode verwendet.[245] Die Treffergenauigkeit ist mit ca. 75 % sehr hoch, die Vollständigkeit mit nur 240 Termkandidaten aus 25.000 Wörtern aber äußerst gering. In einem zweiten Versuch wurde eine hybrid linguistisch-statistische Methode angewandt. Dafür musste zunächst ein Wörterbuch erstellt und ein POS-Tagger trainiert werden. Mit 819 Termkandidaten war die Ausbeute höher, es fehlen aber Angaben über die Treffergenauigkeit.

Hong et al.[246] extrahieren mit Hilfe von manuell erstellten linguistischen Regeln, mit einem statistischen Filter und zusätzlich einer manuellen Stoppwortliste aus einem Text mit 74676 Wörtern 3176 Termkandidaten, von denen in der ex-post-Bewertung nur 2372, also 75 % Terme waren. Auf den obigen Versuch umgerechnet bedeutet das, dass etwa 12 Termkandidaten extrahiert würden und nur neun Terme wären. Das Ergebnis der Extraktion von Hong et al. war also trotz manueller Bearbeitung nicht besser, so weit man die Versuche für vergleichbar hält.[247] Die Ergebnisse sind also vergleichbar gut wie diejenigen von Daille et al.[248]

An diesen Versuchen zeigt sich, auf welche Schwierigkeiten die Termextraktion mit nichteuropäischen Sprachen stößt, zeigt aber zugleich Möglichkeiten zur Integration von linguistischen Ansätzen. Es darf vermutet werden, dass die meisten Terminologieprojekte für Minderheitensprachen von Vornherein auf eine Termextraktion verzichten und traditionell arbeiten.

---

[242] Ahmad K., Davies A. E. (1994), Weirdness in Special-language Text: Welsh Radioactive Chemicals Texts as an Exemplar, S. 22-52 in: Veröffentlichungsreihe des Internationalen Instituts für Terminologieforschung (IITF-Series) 1994, Band 5, Nr. 2, TermNet Wien 1994.

[243] Daille B., Enguehard C., Jacquin C., Raharinirina R. L., Ralalaoherivony B. S., Lehmann C. (2000), Traitement automatique de la terminologie en langue malgache, S. 225–242 in: Mariani J., Masson N., Néel F., Chibout K. (Hrsgg.) (2000), Ressources et évaluation en ingénierie des langues, Actualités scientifiques - Universités Francophones, De Boek and Larcier Paris 2000.

[244] Madagassisch gehört zum westindonesischen Zweig der austronesischen Sprachfamilie. Madagassisch ist Haupt- und Nationalsprache in Madagaskar und wird von 12 Millionen Menschen gesprochen.

[245] Enguehard C., Pantera L. (1994): Automatic natural acquisition of a terminology, S. 27–32 in: Köhler R. (Hrsg.), Journal of Quantitative Linguistics 2(1), International Quantitative Linguistics Association (IQLA) Trier 1994.

[246] Hong M., Fissaha S., Haller J. (2001), Hybrid Filtering for Extraction of Term Candidates from German Technical Texts, Poster in: Tagungsband Terminology and Artificial Intelligence (TIA) 2001, Institut National de l'Information Scientifique et technique (INIST) Nancy 2001, http://www.iai.uni-sb.de/docs/term_extract.pdf : 5.10.2004.

[247] Der Vergleich von Termextraktionsversuchen ist schwierig, weil in der Evaluierung meist kein oder nur ein geschätzter *recall* angegeben wird, da die Gesamtanzahl der Terme im Text nicht bekannt ist.

[248] Daille B. et al. (2000) a.a.O.

## 10. Bewertung der Experimente zur Termextraktion

Die Ergebnisse zeigen, dass die Termextraktion durch Beispielterme eine Alternative zu den statistischen und linguistischen Methoden darstellt. Das ist vor allem dann interessant, wenn die statistischen oder linguistischen Voraussetzungen fehlen wie bei Minderheitensprachen oder beim erstmaligen Aufbau von Terminologie. Einige kopierte oder ausgedachte Beispiele genügen, um Muster zu finden, die fast alle Terme eines Textes extrahieren können. Die Eingabetexte können im Unterschied zum statistischen Ansatz beliebig kurz sein. Sie müssen nicht wie beim linguistischen Ansatz durchanalysiert werden. Der Nachteil dieser Anspruchslosigkeit des beispielbasierten Ansatzes ist, dass Wortkombinationen, die oberflächlich Ähnlichkeiten zu Termen aufweisen, als Termkandidaten ausgewählt werden. Das empfiehlt diesen Ansatz für eine Kombination mit anderen, die andere Stärken und Schwächen haben. Auch in der Praxis ist solch eine Kombination möglich, wenn die durch Beispielterme gefundenen Termkandidaten mit einem Hintergrundkorpus oder mit linguistischen Hilfsmitteln weiter gefiltert werden.

Ein interessantes Ergebnis ist, dass die Termextraktion von Rechtsbegriffen nur rechtliches Untersuchungsmaterial erforderte, aber kein rechtliches Trainingsmaterial, denn Fachbegriffe waren nicht wesentlich bessere Beispiele als Schlagwörter eines Allgemeinwörterbuchs. Das könnte darauf hindeuten, dass entweder die morphologische Spezifität von Rechtsbegriffen gering ist oder dass in Rechtstexten kaum nichtrechtliche Begriffe auftauchen.

Eine Idee zur Verbesserung der TE wäre eine Lernschleife. Man würde eine Variante mit hohem *recall* implementieren und wenn die Terminografen auf die wirklichen Terme in der Ergebnisliste zugreifen und weiterbearbeiten, dann könnte diese Terme den Beispielen hinzugefügt werden. Dadurch könnte die Präzision langsam erhöht werden.

## 11. Überlegungen zum Einsatz der TE zur Textindexierung

Kann die Termextraktion durch Beispielterme auch zur Textindexierung verwendet werden? Mit welcher Technik können beispielsweise (Korpus-)Texte mit automatisch erkannten Rechtsbegriffen indiziert werden?

Ein Index ist vergleichbar mit einem Schlagwortverzeichnis in einem Buch: Auf eine Anfrage wird im Verzeichnis nachgeschlagen und so die relevante Stelle gefunden. Voraussetzung dafür ist, dass im Index diejenigen Wörter stehen, die als unterscheidungsrelevant gesucht werden.

Damit automatisch extrahierte Termkandidaten einen guten Index darstellen, müssten Termkandidaten gleichzeitig geeignete Indexterme sein, also den Text geeignet repräsentieren. Um darüber Aussagen machen zu können, muss ein Maß für die ‚Geeignetheit' angewandt werden. Welche Terme geeignet sind, hängt aber stark von der Textsammlung und dem Informationssuchenden ab (eingehende Diskussion im Kapitel zur Wissensorganisation). Ergebnisse wären in diesem Bereich ebenso aufwändig wie schlecht übertragbar, so dass die Hypothese im Folgenden theoretisch untersucht wird.

### 11.1. Einwände gegen Termkandidaten als Indexterme

Die gewichtigsten Einwände gegen die Verwendung von automatisch extrahierten Termkandidaten als Indexterme sind:

1. Die Indexterme repräsentieren die Dokumente schlecht, wenn Themen nicht ausdrücklich bezeichnet werden und wenn ausdrückliche Bezeichnungen nicht dem Thema entsprechen.

Der erste Teil der Kritik beruht auf der Variabilität der Lexikalisierung von Konzepten. Ein Konzept kann tatsächlich auf verschiedene Weise angesprochen werden, aber auf irgendeine Weise wird es angesprochen werden müssen. Die extrahierten Begriffe müssen ebenso subtil interpretiert werden wie sie im Ausgangstext angesprochen werden. Außerdem kann gerade die Lexikalisierung entscheidender Anhaltspunkt für die Anfrage sein. Beispielsweise könnte nach „Fuchs+Trauben" gesucht werden, um Äsops Fabel[249] zu finden, auch wenn weder *vulpes* (Fuchs) noch *vitis vinifera* (Weinstock) mit dem Sinn der Fabel zu tun haben.

Damit trifft dieser Einwand eher Ontologien, die gerade die Sinnrelationen herausarbeiten wollen und Thesauri, die mit kontrolliertem Vokabular arbeiten, und in geringerem Maße Ansätze, die von den vorhandenen Zeichenketten im Text ausgehen.

Der zweite Teil des Einwands ist schwerwiegender. Es besteht die vage Hoffnung, dass die Begriffe, die nicht Thema sind, vor allem in Sachtexten weniger spezifisch und weniger frequent sind als die thematisierten. Letztlich bleibt dies aber ein grundsätzliches Problem aller Dokumentrepräsentationen, die von Zeichenketten ausgehen.

2. Die vielen falschen Indexterme machen den automatisch erzeugten Index unbrauchbar.

Für kleine Textkollektionen können die relevanten Indexterme durch Fachleute ausgewählt werden. Bei sehr großen Textkollektionen könnte man nur diejenigen Indexterme verwenden, die zumindest zweimal vorkommen und somit Nichtterme ausfiltern. Falsche Indexterme sind nur dann störend, wenn durch sie falsche Treffer entstehen oder relevante Dokumente nicht gefunden werden. Das passiert dann, wenn zwischen der Repräsentation der Retrievalanfrage und der Repräsentation des Textes aufgrund der falschen Indexterme ein anderes, falsches Ähnlichkeitsmaß errechnet wird. Eine geeignete Ähnlichkeitsermittlung sollte das verhindern können. Bezieht man beispielsweise nur volle Treffer in die Ähnlichkeit ein, dann werden irreführende Nichtterme niemals Treffer erzeugen. Andererseits können auch Nichtterme Retrievalqualität haben, wenn kein Indexterm mit voller Übereinstimmung vorliegt, etwa bei Abkürzungen, Eingabefehlern und Schreibvarianten (z.B. „Rechtssprechung" statt „Rechtsprechung"). Das Retrieval könnte textcharakteristische Textfetzen als Mittelweg zwischen Index und Volltext benutzen (im Beispiel: „sprechung").

Je größer die Textkollektion im Verhältnis zu den möglichen Suchanfragen, umso wahrscheinlicher werden Volltreffer in den Indextermen vorliegen und umso weniger fallen falsche Indexterme ins Gewicht.

Grundsätzlich stellt sich jedoch die Frage, ob über die Retrievalfunktion hinaus weitere Zwecke mit dem Index verfolgt werden. Für Retrievalzwecke genügen „terminologisch relevante Kollokationen",[250] die aber nicht unbedingt konzeptbeschreibende Terme sind. Ein Index mit vielen Nichttermen ist für Menschen unverständlich und weder manuelle Klassifikation noch die Themenauswahl direkt aus der Indexliste sind dann möglich. Um zu verhindern, dass Nichtterme indexiert werden, müsste die Ergebnisliste rechtzeitig, etwa nach 10 Termen, abgeschnitten werden. Es wäre noch zu untersuchen, wie sich diese verkürzte Indexierungsbreite[251] auf die Retrievalqualität aus-

[249] Die Fabel vom Fuchs und den Trauben (nach Äsop): Ein Fuchs ging an einer Mauer entlang. Oben stand ein Weinstock, von dem blaue Trauben herabhingen. Der Fuchs sprang in die Höhe, um zu den Trauben zu gelangen. Aber er konnte sie nicht erreichen. Da ging er weiter seines Weges und sagte: „Die Trauben sind mir zu sauer."

[250] Heid U. (1999), Extracting terminologically relevant collocations from German technical Texts, S. 241-255 in: Sandrini P. (Hrsg.) (1999), Proceedings of the 5th International Congress on Terminology and Knowledge Engineering (TKE) 1999 Innsbruck, TermNet Wien 1999, http://citeseer.nj.nec.com/heid99extracting.html : 5.10.2004.

[251] Indexierungsbreite ist das Ausmaß, in dem der fachliche Inhalt eines Dokuments von seinen Indexbegriffen abgedeckt wird. Um die Indexierungsbreite in ein Maß fassen zu können, wird davon ausgegangen, dass die Abdeckung proportional mit der Anzahl der Indextermini steigt. Man gibt daher die durchschnittliche Anzahl der Indexbegriffe pro Dokument an. Knorz G. (1997), Indexieren, Klassieren, Extrahieren, S. 120-140 in: Buder M., Rehfeld W., Seeger T.,

wirkt, denn je feiner die Dokumente unterschieden werden sollen, umso mehr Indexterme sind nötig. Zumindest für große Textkollektionen muss auf weitere Termkandidaten zurückgegriffen werden.[252]

3. Die Indexterme spannen einen *n*-dimensionalen Raum auf, obwohl Terme mit semantisch unerheblichen Abweichungen (Mehrzahl, Schreibvarianten) in derselben Dimension liegen.

Eine Lösungsmöglichkeit wäre die weitere Bearbeitung der Indexliste, indem Terme mit geringer grafemischer Abweichung unter dem häufigsten Begriff zusammengefasst werden oder alle als Treffer gewertet werden.[253] Dies funktioniert nur bei Abweichungen in wenigen Zeichen und löst nicht das Problem bei Ellipsen und Synonymen.

Bei einer Klassifikationsaufgabe fällt dieses Problem viel weniger ins Gewicht, weil die Ähnlichkeit nicht über einige wenige Terme, sondern über alle Wörter berechnet wird. Je weniger Indexterme also verwendet werden, umso drängender wird das Problem. Es besteht daher ein Zielkonflikt zwischen der Extraktion von möglichst vielen Termen und möglichst präzisen Termen (Einwand 2).

4. Lange Texte produzieren mehr Indexterme und ihre Relevanz wird überbewertet.

Eine Lösung wäre es, das Maß der Ähnlichkeit durch die Übereinstimmung in Buchstaben zu berechnen und den Wert des Treffers mit der Anzahl der vorhandenen *n*-Gramme ins Verhältnis zu setzen. Die überbewerteten langen Texte würden damit zumindest an das Ende der Ergebnisliste gedrückt. Außerdem haben lange Texte das Handicap, dass mehrfach vorkommende Begriffe nur einmal im Index auftauchen. Das wirkt der ungerechtfertigten Höherbewertung sich stark wiederholender Texte entgegen.

Nach dem Zipfschen Gesetz müssten sich die zusätzlich hinzukommenden Begriffe und die Begriffswiederholungen bei Einberechnung der Textlänge ins Ähnlichkeitsmaß die Waage halten.

11.2. Vorteile der Termextraktion beim Indexieren

Der Vorteil der beispielbasierten Indexierung liegt sicherlich in den **geringen sprachlichen und logistischen Voraussetzungen**, die diese Methode stellt. Es genügen zwanzig Texte und hundert Beispielsbegriffe in einer Sprache, um einen Index aufzubauen. Damit eignet sich die Methode besonders für Minderheitensprachen. Es muss sich nicht einmal um eine Sprache handeln, denn die Methode funktioniert auch bei anderen Zeichenketten. Es können also auch Musiknoten, Bilder und überhaupt **alle digitalen Dokumente** indexiert werden, sofern man Indexierungsbeispiele beibringen kann und nötigenfalls einige einfache allgemeine Regeln angibt.

Der **Aufbau eines Indexes nach der Struktur der vorgegebenen Beispiele** bringt weitere Vorteile mit sich. Paijmans[254] zeigt, dass weder durch eine bestimmte Stellung von Wörtern im Text noch durch ihre Nähe zu Signalwörtern (*cuewords*) ein höherer Informationsgehalt festgestellt wer-

---

Strauch D. (Hrsgg.) (1997): Grundlagen der praktischen Information und Dokumentation, Saur Verlag München u.a., 4. Aufl. 1997.

[252] Eine Möglichkeit, den Index weiterhin menschenverständlich zu halten wäre der Abgleich mit einem Thesaurus bzw. einer umfangreichen Liste zugelassener Indexterme.

[253] Ekmekçioglu F. Ç., Lynch M.F., Robertson A.M., Sembok T.M.T., Willett P. (1996), Comparison of n-gram matching and stemming for term conflation in English, Malay, and Turkish texts, S. 1-14 in: Text Technology 6/1996, sowie in: Information Research, Vol. 2 No. 2, Oktober 1996 http://informationr.net/ir/2-2/paper13.html : 5.10.2004.

[254] Paijmans H. (1997), Gravity Wells of Meaning: detecting Information-Rich passages in Scientific Texts, S. 520-536 in: Journal of Documentation 53(5), 1997.

den konnte. Paijmans findet aber eine größere Bedeutung von bestimmten Wortarten, etwa von Adjektiven, Verben und Substantiven. Diese Erkenntnis spricht für die beispielbasierte Indexierung, die bestimmte Wortarten bevorzugt. Durch die Modellgenerierung anhand der Beispiele werden nur beispielähnliche Zeichenketten indexiert, also bei Eingabe von Adjektiven praktisch nur Adjektive, sofern die Ähnlichkeitsparameter gekonnt gewählt werden.

Aufgrund dieser **Flexibilität** und des geringen Aufwands bei der Neuerstellung eines Indexes kann zum Import und Export von Dokumenten stets ein neuer, gemeinsamer Index erstellt werden. Das besonders aufwändige Übertragen von Indexkategorien entfällt.

Die Flexibilität der Indexierung kann ausgenutzt werden, um die gleiche Dokumentensammlung gleichzeitig für verschiedene Zwecke verschieden zu indexieren. Eine medizinische Datenbank könnte beispielsweise für Fachleute aufgrund lateinischer Fachbegriffsbeispiele und für Lerner der medizinischen Fachsprache aufgrund von Allgemeinwörterbuchbeispielen erstellt werden. Aufgrund der Beispiele werden jeweils die stärker informationstragenden Terme extrahiert und indiziert.

Nebeneinander bestehende Indexe wären auch für Dokumente mit besonderen Textstrukturen und -eigenschaften geeignet. Dann könnten automatisch verschiedene Indexe für die Zusammenfassungen, den Text und die Bibliographie von wissenschaftlichen Artikeln zur Repräsentierung verwendet werden.

Durch die **Veränderung des Termfrequenzparameters** kann für eine bestimmte Dokumentkollektion oder überhaupt für eine Sprache indexiert werden. Wird für die Gewichtung eines Termkandidaten seine Frequenz im Dokument mit seiner Frequenz in der Dokumentsammlung ins Verhältnis gesetzt, so beschreibt das Maß die Charakteristik eines Dokuments im Gegensatz zu anderen Dokumenten der Sammlung.[255] Wird die Dokumentfrequenz mit einem Allgemeinkorpus ins Verhältnis gesetzt, so bleibt die Spezifizität (z.B. die Fachsprachlichkeit) eines Dokuments auch dann erhalten, wenn es sich in einem Fachkorpus befindet. Die Folge wäre einerseits, dass sich die verschiedenen fachsprachlichen und die allgemeinsprachlichen Dokumente voneinander trennen, andererseits aber, dass viele Dokumente dieselben Indexterme tragen und Suchanfragen stets extrem viele oder extrem wenige Dokumente ergeben. Die Gewichtung sollte also danach gewählt werden, ob die Relevanz der Ergebnisdokumente für eine Suchanfrage im Kontext idealer Information beurteilt wird (absolut) oder ob die Relevanz im Kontext der Dokumentenkollektion (relativ) beurteilt wird.

Schließlich kann die beispielbasierte Indexierung mit statistischen oder regelbasierten Methoden verbunden und verbessert werden, wenn die nötigen Ressourcen vorhanden sind. Das könnte insbesondere bei Sprachen nötig sein, bei denen die Ergebnisse aufgrund vielfältiger Möglichkeiten zur Zusammensetzung, Beugung und Ableitung von Wörtern nicht hinreichend gut sind. Obwohl der Ansatz unabhängig von der Eingabesprache funktioniert, funktioniert er nicht stets gleich gut. Die Extraktion/Indexierung wird bei romanischen Sprachen besser sein als bei germanischen oder slawischen Sprachen.

### 11.3. Ausblick

Interessant wäre, ob sich die Retrievalqualität steigern ließe, wenn man nicht Wörterbucheinträge oder Fachbegriffe zugrundelegt, sondern tatsächliche Suchanfragen. Zwar kann erst gesucht werden, wenn bereits ein Index aufgebaut ist, aber der Index könnte später mit gespeicherten Suchanfragen noch einmal neu aufgebaut werden. Einerseits ist zu vermuten, dass die Ähnlichkeit zwischen Anfrage und Indexbegriffen damit größer wird. Andererseits werden Suchanfragen wohl häufig als Kombinationssuche nach mehreren Begriffen formuliert, die Termextraktion würde die kombinierten Worte aber fälschlicherweise als zusammengehörend interpretieren und unsinnige Muster generieren.

---

[255] Oder genauer: Die Charakteristik des Termkandidaten für dieses Dokument im Gegensatz zu seiner Charakteristik für den Durchschnitt aller anderen Dokumente der Sammlung.

Die Ausnutzung bestehender Termextraktionsmethoden zum Indexieren für Retrievalzwecke erscheint insgesamt als möglicher Ansatz, muss ihre Tauglichkeit aber erst noch experimentell erweisen. Für Versuche steht die Termextraktion unter der Adresse http://dev.eurac.edu:8080/cgi-bin/index/TermExtract für die Sprachen Deutsch, Englisch, Französisch, Italienisch, Ladinisch allgemein, Grödnerisch, Gadertalerisch und Fassanisch bei automatischer Sprachenidentifizierung zur Verfügung. Die Texte zur Indexierung/Extraktion können als URL oder über ein Textfenster eingegeben werden. Die Sortierung der Termkandidaten kann nach Termfrequenz, Termfrequenz mal Länge, Ähnlichkeit von *n*-Grammen, *weirdness-ratio* oder *mutual information* in Tabellen oder Absätzen erfolgen. Wenn beim Indexieren nur einige Begriffe verwendet werden sollen, dann kann man die Ergebnisliste nach der gewünschten Anzahl abschneiden, womit nur die besten Termkandidaten verbleiben. Das Programm gibt die Möglichkeit zur Begrenzung auf 5 bis 30.000 Termen. Dann kommt dem *ranking* besonderes Gewicht zu, das im Versuch gute Ergebnisse lieferte.

## 12. Zusammenfassung

Die in diesem Kapitel vorgestellten terminografischen Instrumente setzen voraus, dass Texte vorliegen, die interessante Begriffe enthalten (Termextraktion) und daher aufbewahrenswert sind (indexierte Speicherung der Texte). Diese Tätigkeit mündet zwangsläufig in eine **Dokumentverwaltungsstrategie**, die jenseits des Korpusaufbaus (Kapitel 1) dazu tendiert, das in elektronischen Texten und Termini enthaltene Wissen zu ordnen.

Durch die bisher besprochenen Werkzeuge kann rechtliches Wissen effektiv erlangt, verwaltet und präsentiert werden. Dabei wurde das Wissen nach traditionellen, technischen und linguistischen Gesichtspunkten strukturiert. Die Gebrauchstauglichkeit aller Werkzeuge steht und fällt aber nicht nur damit, ob sie funktionieren, sondern auch damit, ob die Ergebnisse nützlich und sinnvoll sind. Die entscheidenden Fragen hierfür sind: Welche Struktur hat Rechtswissen (abgesehen von Fachgebieten)? Kann man diese Struktur abbilden und lässt sich dadurch vielleicht sogar neues Wissen gewinnen? Wie hängen technische und inhaltliche Strukturen miteinander zusammen? All diese Fragen beziehen sich auf die Ordnung (rechtlichen) Wissens, mit der sich das folgende Kapitel beschäftigen wird.

# Literaturangaben zu Kapitel 4

Ahmad K., Davies A. E. (1994), Weirdness in Special-language Text: Welsh Radioactive Chemicals Texts as an Exemplar, S. 22-52 in: Veröffentlichungsreihe des Internationalen Instituts für Terminologieforschung (IITF-Series) 1994, Band 5, Nr. 2, TermNet Wien 1994.

Arppe A. (1995), Term extraction from unrestricted text, short paper in: Koskenniemi K. (Hrsg.) (1995), Proceedings of the 10th Nordic Conference of Computational Linguistics (Nordiska datalingvistdagarna: NoDaLiDa 1995), Helsinki 1995, http://www.lingsoft.fi/doc/nptool/term-extraction.html : 5.10.2004.

Bourigault D., Jacquemin C. (1999), Term extraction and term clustering. An integrated platform for computer-aided-terminology, S. 15-22 in: Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics (EACL 1999), EACL Bergen 1999, http://citeseer.nj.nec.com/bourigault99term.html : 5.10.2004.

Brekke M., Myking J., Ahmad K. (1996), Terminology management and lesser-used living languages: A critique of the corpus-based approach, S. 179-189 in: Sandrini P. (Hrsg.) (1996), Proceedings of the 4th International Congress on Terminology and Knowledge Engineering (TKE'96) Innsbruck, TermNet Wien 1996, ftp://ftp.ee.surrey.ac.uk/pub/research/AI/TKE.papers/Postscript_versions/Terminology_Management.ps.gz : 5.10.2004.

Cabré Castellvì M.T., Estopà Bagot R., Palatresi J. V. (2001), Automatic term detection: A review of current systems, S. 53-89 in: Bourigault D., Jacquemin C., L'Homme M. C. (Hrsgg.) (2001), Recent Advances in Computional Terminology, Natural Language Processing Band 2, John Benjamins Publishing Company Amsterdam/ Philadelphia 2001.

Church W., Hanks P. (1989), Word association norms, mutual information and lexicography, S. 76–83 in: Proceedings of the 27th Annual Meeting of the Annual Meeting of the Association for Computational Linguistics (ACL), ACL Vancouver 1989, http://citeseer.nj.nec.com/church89word.html : 5.10.2004.

Daille B. (1995), Combined approach for terminology extraction: lexical statistics and linguistic filtering, S. 515-521 in: Proceedings of the 15th International Conference on Computational Linguistics Kyoto (COLING 94) 1994, ICCL Kyoto 1994, http://acl.ldc.upenn.edu/C/C94/C94-1084.pdf : 4.10.2004.

Daille B., Enguehard C., Jacquin C., Raharinirina R. L., Ralalaoherivony B. S., Lehmann C. (2000), Traitement automatique de la terminologie en langue malgache, S. 225–242 in: Mariani J., Masson N., Néel F., Chibout K. (Hrsgg.) (2000), Ressources et évaluation en ingénierie des langues, Actualités scientifiques - Universités Francophones, De Boek and Larcier Paris 2000.

Ekmekçioglu F. Ç., Lynch M. F., Robertson A. M., Sembok T. M. T., Willett P. (1996), Comparison of n-gram matching and stemming for term conflation in English, Malay, and Turkish texts, S. 1-14 in: Text Technology, 6/1996, sowie in: Information Research, Vol. 2 No. 2, Oktober 1996,
http://informationr.net/ir/2-2/paper13.html : 5.10.2004.

Enguehard C., Pantera L. (1994), Automatic natural acquisition of a terminology, S. 27-32 in: Köhler R. (Hrsg.), Journal of Quantitative Linguistics 2(1), International Quantitative Linguistics Association (IQLA), Trier 1994.

Glück H. (2000), Metzler Lexikon Sprache, Stichwort 'Funktionswort' S. 226, Metzler Stuttgart 2000.

Gruber T. R. (1993), Towards Principles for the Design of Ontologies Used for Knowledge Sharing, in Guarino N., Poli R., (Hrsgg.) (1993), Formal Ontology in Conceptual Analysis and Knowledge Representation, Kluwer Academic Publishers Deventer (NL) 1993.

Heid U. (1999), Extracting terminologically relevant collocations from German technical Texts, S. 241-255 in: Sandrini P. (Hrsg.) (1999), Proceedings of the 5th International Congress on Terminology and Knowledge Engineering (TKE) 1999, TermNet Wien 1999,
http://citeseer.nj.nec.com/heid99extracting.html : 5.10.2004.

Hong M., Fissaha S., Haller J. (2001), Hybrid filtering for extraction of term candidates from German technical texts, Poster in: Proceedings of Terminologie et Intelligence Artificielle (TIA), Institut National de l'Information Scientifique et Technique (INIST) Nancy 2001,
http://www.iai.uni-sb.de/docs/term_extract.pdf : 5.10.2004.

ISO 2788, 1986:2

Jacquemin C. (1999), Syntagmatic and paradigmatic representation of term variation, S. 341-348 in: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL) 1999, ACL Maryland 1999,
http://citeseer.nj.nec.com/jacquemin99syntagmatic.html : 5.10.2004.

Jacquemin C., Bourigault D. (2003), Term Extraction and Automatic Indexing, in: Handbook of Computational Linguistics, Mitkov R. (Hrsg.) (2003), Oxford University Press Oxford 2003.

Knorz G. (1997), Indexieren, Klassieren, Extrahieren, S. 120-140 in: Buder M., Rehfeld W., Seeger T., Strauch D. (Hrsgg.) (1997), Grundlagen der praktischen Information und Dokumentation, Saur Verlag München u.a., 4. Aufl. 1997.

Maynard D., Ananiadou S. (2001), Term extraction using a similarity-based approach, S. 53-89 in: Bourigault D., Jacquemin C., L'Homme M. C. (Hrsgg.) (2001), Recent Advances in Computional

Terminology, Natural Language Processing Band 2, John Benjamins Publishing Company Amsterdam/ Philadelphia 2001,
http://citeseer.nj.nec.com/maynard99term.html : 5.10.2004.

Merkel M., Nilsson B., Ahrenberg L. (1994), A phrase-retrieval system based on recurrence, S. 43-56 in: Proceedings of the Second Annual Workshop on Very Large Corpora (WVLC-2), Kyoto 1994,
http://www.ida.liu.se/~magme/publications/kyoto--94.pdf : 13.9.2004.

Merkel M., Mikael A. (2000), Knowledge-lite extraction of multi-word units with language filters and entropy thresholds, S. 737–746 in: Proceedings of the 6th International Conference on Content-Based Multimedia Information Access (RIAO) 2000, Vol. 1, Collège de France Paris 2000,
http://citeseer.nj.nec.com/merkel00knowledgelite.html : 4.10.2004.

Nakagawa H. (2001), Experimental evaluation of ranking and selection methods in term extraction, S. 303-325 in: Bourigault D., Jacquemin C., L'Homme M. C. (Hrsgg.) (2001), Recent Advances in Computional Terminology, Natural Language Processing Band 2, John Benjamins Publishing Company Amsterdam/ Philadelphia 2001,
http://www.r.dl.itc.u-tokyo.ac.jp/~nakagawa/academic-res/termrec.pdf : 5.10.2004.

Paijmans H. (1997), Gravity Wells of Meaning: detecting Information-Rich passages in Scientific Texts, S. 520-536 in: Journal of Documentation 53(5), 1997,
http://pi0959.kub.nl/Paai/Onderw/V-I/Content/grav_wells.html : 5.10.2004.

Quasthoff U., Biemann C., Wolff C. (2002), Named Entity Learning and Verification: Expectation Maximization in Large Corpora, S. 8-14 in: Roth D., Bosch A. van den (Hrsgg.) (2002), Proceedings of the 6th Workshop on Computational Language Learning (CoNLL), 2002 Taipei.

Smadja F. (1993), Retrieving collocations from text: Xtract, S. 142-176 in: Computational Linguistics 19 (1),
http://acl.ldc.upenn.edu/J/J93/J93-1007.pdf: 4.10.2004.

Soininen P., Voutilainen A., Tapanainen P. (1999), An experiment in automatic term extraction, S. 234–240 in: Sandrini P. (Hrsg.) (1999), Proceedings of the 5th International Congress on Terminology and Knowledge Engineering (TKE) 1999, TermNet Innsbruck 1999.

Somers H. (2003), Machine Translation: Latest Developments, S. 512-528 in: Mitkov R. (Hrsg.) (2003), The Oxford Handbook of computational Linguistics, Oxford University Press Oxford 2003.

Streiter O., De Luca E. W. (2003), Example-based NLP for Minority Languages: Tasks, Resources and Tools, S. 233-242 in: Proceedings of the Conférence sur le Traitement Automatique des Langues Naturelles (TALN), Association pour le Traitement Automatique des Langues (ATALA) Batz-sur-Mer 2003,

http://dev.eurac.edu:8080/taln/accepted/streiter_de_luca.pdf : 5.10.2004.

Streiter O., Hsueh P. (2000), A case-study on example-based parsing, in: Tagungsband der International Conference on Chinese Language Computing (ICCLC) Chicago, 2000.
http://citeseer.nj.nec.com/streiter00case.html : 23.12.2003.

Streiter O., Zielinski D., Ties I., Voltmer L. (2002), Example-based Term Extraction for Minority Languages: A case-study on Ladin, in: Tagungsband von Soziolinguistica Y Language Planning, SPELL St. Ulrich (Italien) in Vorbereitung,
http://dev.eurac.edu:8080/autoren/publs/termex5.pdf : 5.10.2004.

Streiter O., Zielinski D., Ties I., Voltmer L. (2003), Term Extraction for Ladin: An Example-based Approach, S. 275-284 in: Tagungsband der Konferenz zum Traitement Automatique des Langues Naturels (TALN) 2003, Batz-sur-Mer, 2003,
http://dev.eurac.edu:8080/autoren/publs/taln_minority_streiter_et_all.pdf : 5.10.2004.

Van Rijsbergen C. J. (1983), Information Retrieval, Butterworths London 1983.

Zielinski D. (2002), Computergestützte Termextraktion aus technischen Texten, Diplomarbeit Universität des Saarlands, Saarbrücken 2002,
http://www.iai.uni-sb.de/~mt-dept/texte/zielinski.pdf: 4.10.2004.

Internetquellen in Zitierreihenfolge:
http://www.ifi.unizh.ch/CL/InteractiveCLtools/index.php : 4.10.2004.
http://www.d.umn.edu/~tpederse/nsp.html : 13.9.2004.
http://dev.eurac.edu:8080/cgi-bin/index/TermExtract : 5.10.2004.

Das Termextraktionsprogramm kann unter der Adresse http://dev.eurac.edu:8080/perl/all.tar.gz heruntergeladen werden. Es läuft unter der graphischen Oberfläche von BISTRO auf der Webseite http://dev.eurac.edu:8080/cgi-bin/index/TermExtract.

Kapitel 5

# V. Organisation rechtlichen Wissens

## Überblick über Wissensorganisationssysteme (insbesondere Ontologien) für Recht

Dieses Kapitel gibt zunächst einen Überblick über die vielfältigen Möglichkeiten, Methoden und Ziele der Einteilung und Repräsentation von (rechtlichem) Wissen. Die verschiedenen Wissensorganisationssysteme werden genannt und durch ihre Funktion, ihren Zweck sowie ihr klassisches Anwendungsgebiet beschrieben. Eine besonders hoch entwickelte Methode der Wissensrepräsentation mit formaler Sprache, die logische Operationen ermöglicht, ist eine Ontologie. Die verschiedenen Ansätze zur Einteilung rechtlichen Wissens mit Ontologien werden vorgestellt und danach eingeteilt, wie weitgehend sie Wissen formalisieren und inwieweit durch die Anwendung formaler Logik logische Schlüsse aus der Ontologie gezogen werden können. Anschließend wird anhand einer Datenbank mit Rechtstexten vorgeführt, wie eine Ontologie erstellt werden kann.

### 1. Begriffsklärungen und Einführung

Die folgenden Definitionen sind Arbeitsdefinitionen, da sich für keinen der Begriffe eine einheitliche Definition durchgesetzt hat.

**Wissen** soll hier im Sinn der Informationstheorie[256] verstanden werden. Das bedeutet, dass Zeichen verwendet werden, die durch eine Syntax zu Daten verbunden werden können. Daten, die eine Bedeutung tragen, sind Information. Durch die Verknüpfung von Information, z.B. durch Kontext oder Erfahrung, entsteht Wissen.

„**Wissensrepräsentation** bezeichnet einerseits die Realisierung abstrakten Wissens in einem konkreten (physikalischen) System […] und andererseits die Strukturen, die sich aus der Interaktion eines vorstrukturierten informationsverarbeitenden Systems mit seiner Umwelt […] ergeben."[257] Die Wissensrepräsentation ist eine Voraussetzung für Künstliche Intelligenz. Sie ist auf die symbolischen Methoden der Computerlinguistik angewiesen, und konkret auf Wissensrepräsentationsformalismen zur Modellierung von Wissen. Sie stützen sich auf die Theorie der Automaten und formalen Sprachen, die Graphentheorie und die Logik.[258]

**Wissensorganisation** ist die Strukturierung von Information zum Speichern, Wiederfinden, Verwalten, Erschließen und allgemein zur optimalen Nutzung von Wissen. Diese Bezeichnung soll Verwechslungen vorbeugen, insbesondere mit dem viel verwendeten Begriff Klassifikation. Eine (Wissens-) **Klassifikation** soll hier ausschließlich eine monohierarchische (keine Klasse hat mehr

[256] Informationstheorie und Kommunikationstheorie (*information and communication theory*) von Shannon C. E. (1998), The Mathematical Theory of Communication, University of Illinois Press Urbana/ Chicago 1998.

[257] Carstensen K.-U. (2004), Nicht-sprachliches Wissen, S. 448-454 in: Carstensen K.-U., Ebert C., Endriss C., Jekat S., Klabunde R., Langer H. (Hrsgg.) (2004), Computerlinguistik und Sprachtechnologie – Eine Einführung, Spektrum Verlag Heidelberg 2004, S. 449.

[258] Rumpf C., Automatentheorie, formale Sprachen und Computerlinguistik. In Seminar Automaten, Androiden und KI, Heinrich-Heine-Universität Düsseldorf, 10. Januar 2001, http://www.phil-fak.uni-duesseldorf.de/~rumpf/talks/automatentheorie.pdf: 10.11.2003.

als einen Oberbegriff) und analytische (jede Klasse unterscheidet sich durch mindestens ein Merkmal von anderen) Einteilung (z.B. ein Thesaurus oder eine Ontologie) bezeichnen.

Ein **Wissensorganisationssystem** ist der hier gebrauchte Überbegriff für jegliche Ordnung von Wissen.

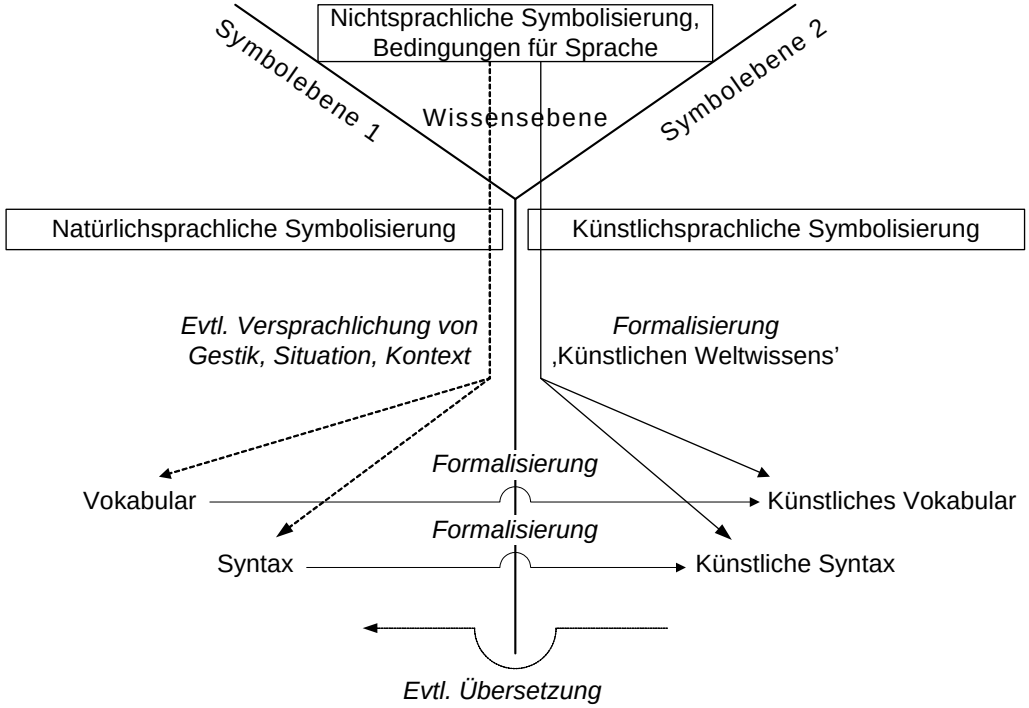
In der Einleitung wurde dargelegt, welch enorm wichtige Rolle die Schrift für die Speicherung und Weitergabe von rechtlichem Wissen spielt. Bis vor kurzem war die Schrift an einen physischen Datenträger (z.B. ein Buch) gebunden, so dass die Ordnung der Sachen (der Bücher) eine Ordnung des repräsentierten Inhalts mit sich brachte. Die heutige Bibliotheks- und Dokumentationswissenschaft baut auf jahrtausendealte Erfahrungen im Ordnen nach Themen und Inhalten auf, für die eine Vielzahl von geeigneten Ordnungsschemata, Begriffssysteme und Dokumentationsregeln erfunden und erprobt wurden. Die Loslösung der Information von materiellen Datenträgern hat dreierlei bewirkt. Erstens müssen zu speichernde Informationen praktisch nicht mehr ausgewählt werden, weil der zur Verfügung stehende Speicherplatz gegen Unendlich tendiert. Zweitens entfällt mit der Entkoppelung von Information und Medium (z.B. Schreibtafeln oder Bücher) die auf physikalischen Trägern stets implizit gegebene Ordnung. Daraus folgt drittens, dass die einzelnen Informationsträger statt monohierarchisch auch in eine andere Ordnungsstruktur gebracht werden kann (z.B. polyhierarchisch oder assoziativ).[259] Aus diesen drei Gründen kann man sich elektronisch gespeichertes Wissen wie ein Atommodell vorstellen: Eine begrenzte Anzahl kleinster Informationsteile sind miteinander vernetzt und ergänzen sich zu komplexen Systemen.

Die Strukturierung von Information muss nun von der Ordnung und Erschließung einer begrenzten Anzahl materieller, chronologischer und vorsegmentierter Informationseinheiten (z.B. Bücher) auf elektronisches Wissen (z.B. Hypertext, Daten in relationalen Datenbanken) umgestellt werden. Nur wenn thematisch relevante elektronische Informationen ebenso gut gefunden werden können wie traditionell gespeicherte Informationen werden (wissenschaftliche) Erkenntnisse Teil des (Wissenschafts-)Diskurses und seiner Qualitätskontrolle.[260] Die Informatik hat zwar eine Vielzahl von automatischen Techniken entwickelt, um Daten zu speichern, zu repräsentieren und zu verarbeiten, die direkte Suche nach Themen und Inhalten gelingt jedoch noch nicht so gut wie in traditionellen Wissensordnungssystemen. Beispielsweise kann im Internet in Milliarden Dokumenten nach einer Zeichenfolge gesucht werden, aber eine Suche nach Themen scheitert daran, dass die Themen der Dokumente nicht einheitlich und meist gar nicht annotiert sind. Grafik 35 soll die Organisation von Wissen veranschaulichen.

[259] Es gibt natürlich weitere Folgen wie z.B. die zwangsläufig technische Speicherung und Übermittlung elektronischer Information. Das hat zur Folge, dass Information weiterhin kostspielig bleibt (technische Hilfsmittel nötig) und vor allem, dass elektronische Information auch leichter unzugänglich wird (Datenträger und -formate veralten schnell).

[260] Je früher eine gewonnene Erkenntnis anderen Wissenschaftlern zugänglich ist, umso weniger Doppelarbeit gibt es und um so früher kann die ‚Wissensrendite' eingefahren bzw. bei Falsifizierung ‚abgeschrieben' werden. Umstätter betont die Rolle der Informationswissenschaften bei der Rationalisierung der Produktion von Wissen durch Arbeitsteilung. Umstätter W., Die Nutzung des Internets zur Fließbandproduktion von Wissen, S. 179-199 in: Fuchs-Kittowski K., Parthey H., Umstätter W., Wagner-Döbler R. (Hrsgg.), Organisationsinformatik und Digitale Bibliothek in der Wissenschaft, Wissenschaftsforschung Jahrbuch 2000, Gesellschaft für Wissenschaftsforschung Berlin 2001, S. 19-20.

**Grafik 35: Beziehungen von Wissen, natürlicher und künstlicher Sprache**




Diese Grafik zeigt die Trennung des Wissens (Mitte oben) von zwei Symbolebenen der Wissensrepräsentation (links und rechts darunter). Wissen wird einerseits natürlichsprachlich symbolisiert (links) und dadurch kommuniziert. An der Kommunikation von Wissen sind auch die Bedingungen der Kommunikation wie etwa der Kontext und außerdem andere Formen der Kommunikation wie Gestik beteiligt, sie brauchen normalerweise aber nicht versprachlicht zu werden.

Wenn das natürlichsprachlich kommunizierte Wissen nicht wieder durch eine (andere) natürliche Sprache gespeichert, verarbeitet oder wiedergegeben werden soll, dann muss man es durch Formalisierung ‚umsymbolisieren'. Die Symbolisierung in eine künstliche Sprache oder Kunstsprache (englisch: *artificial language*) soll Wissen in eindeutiger Weise repräsentieren. Das ist mit natürlichen Sprachen prinzipiell nicht möglich, daher muss die Kunstsprache von ihrer Konzeption her eine sprachunabhängige Repräsentation sein, die mit formallogischen Verfahren verarbeitet werden kann. Die Kunstsprache repräsentiert Wissen also mittels einer logischen Sprache, mit der Computer etwas anzufangen wissen.[261]

Wenn Menschen auf das in der Kunstsprache symbolisierte Wissen zugreifen möchten, dann ist normalerweise eine (Rück-)Übersetzung in eine natürliche Sprache nötig (siehe Grafik Nr. 1 ganz unten). Manche Menschen können die künstlichsprachliche Symbolisierung direkt verstehen und brauchen keine Übersetzung der Wissensrepräsentation in eine natürliche Sprache. Bibliothekare verstehen Notationen, Musiker Noten, Informatiker Programmcodes und Chemiker Formeln ganz unmittelbar. Für alle anderen kann die künstliche Sprache wieder in eine der natürlichen Sprachen übersetzt werden.

In aller Regel sind künstliche Symbolisierungssprachen wie natürliche Sprachen aufgebaut, also aus Vokabular und Syntax. Das Vokabular ist die Summe aller Bezeichnungen und wird in der Terminologie als Eintrag, in der Mathematik als Element, in der Musik als Note und in der Informatik als Ressource, Knoten oder Entität bezeichnet. Die Syntax ist die Summe aller Verknüpfungsre-

[261] Pansegrau P., Weingarten R. (2000), Metzler Lexikon Sprache, Stichwort 'Künstliche Sprache' S. 390-391, Metzler Stuttgart 2000. Künstliche Sprachen werden nach ihrem Zweck in *artistic languages* zu künstlerischen Zwecken, *auxiliary languages* zur internationalen Kommunikation und *logical languages* für logisch-philosophische Zwecke eingeteilt. In der Wissensorganisation interessiert der logische, bei mehrsprachigem Wissen auch der kommunikative Charakter.

geln für das Vokabular. In der Informatik heißen die definierten Verknüpfungen von Bezeichnungen auch Relationen.

Wird nur das Vokabular formalisiert, kann nur in einer ungeordneten Vokabelsammlung gesucht werden, etwa bei der Suche in kontrollierten Schlagwörtern, lexikalischen Ketten oder einer Normdatenbank. Ein Beispiel für Syntaxformalismen sind die Verknüpfungsregeln der Mengenlehre (‚ist Teilmenge von', ‚ist Schnittmenge von').

Damit die Kommunikation mit künstlicher Sprache präzise genug ist, um automatisch verarbeitet werden zu können, muss auch Wissen formalisiert werden, das in natürlicher Sprache nicht explizit ausgedrückt wird. Es handelt sich dabei um nichtsprachliche Symbolisierungen wie Gestik, um Vorbedingungen der Sprache wie Situation und Kontext (Entschlüsselung deiktischer Kommunikation: „dieses dort drüben"), sowie um aller Kommunikation zugrunde liegende Grundannahmen oder Paradigmen. Mit letzterem ist erfahrenes Vorauswissen gemeint wie etwa, dass Wasser nass ist und die Sonne warm.[262] Dieses implizite Wissen kann entweder explizit versprachlicht („dieses dort drüben": In Richtung des Fingerzeigs steht ein Glas Wasser.) und anschließend formalisiert werden,[263] oder man definiert es direkt in der künstlichen Sprache (Definition: $0^0 = 1$). In der Grafik 35 deuten das die Pfeile von der Wissensebene in die Symbolebenen und von der ersten in die zweite Symbolebene an.

Texte verwenden Schrift zur Repräsentation von Wissen. Der direkteste Zugang zu diesem schriftlichen Wissen ist die natürliche Sprachverarbeitung (*natural language processing* - NLP). Durch *Text Mining* (ein Teilbereich der Computerlinguistik) versucht man, das repräsentierte Wissen automatisch zu extrahieren[264]. Man stößt dabei auf alle traditionellen Problematiken der Wissensorganisation. Vor allem stellt sich die Frage, wie man das Wissen, das man aus seiner natürlichsprachlichen Repräsentation ‚befreit' hat, auf geschicktere Weise symbolisiert. Die Antwort hängt von dem Ziel ab, mit dem man die natürlichsprachliche Symbolebene durch eine künstlichsprachige Symbolebene ersetzt.

Man kann nach den unterschiedlichen Zielen drei Arten Begriffs-, Konzept- und Ordnungssysteme unterscheiden: Terminologien (semasiologisches Nachschlagen von formal beschriebenen Fakten), Dokumentation (onomasiologisches Beschreiben und Wiederauffinden) sowie Datenbanken (automatische Verarbeitung größerer Datenmengen in komplexeren Strukturen).[265] Auf die automatische Verarbeitung künstlicher Sprachen stützt sich die Zielvorstellung des *reasoning*[266] oder der Künstlichen Intelligenz (KI).

---

[262] Der nichtsprachliche Kontext in einem wissenschaftlichen Fachgebiet sind unausgesprochene Paradigmen oder Grundannahmen. Zu den Grundannahmen vergl. Voß a.a.O. S. 7-8 aufbauend auf Thomas Kuhn.

[263] Die weiter wachsende Wissensdatenbank Cyc (von encyclopedia) enthält z.Z. Millionen von expliziten Aussagen des Allgemeinwissens über mehr als 200.000 Begriffe. http://www.cyc.com/cyc/technology/whatiscyc_dir/whatsincyc : 15.3.2004.

[264] Auch *text data mining* genannt. Das Ziel ist, „to discover or derive new information from data, finding new patterns across data sets, and/or separaring signal from noise." Hearst M. A. (2003), Text Data Mining, S. 616-628 in: Mitkov R. (Hrsg.) (2003), The Oxford Handbook of Computational Linguistics, Oxford University Press 2003.

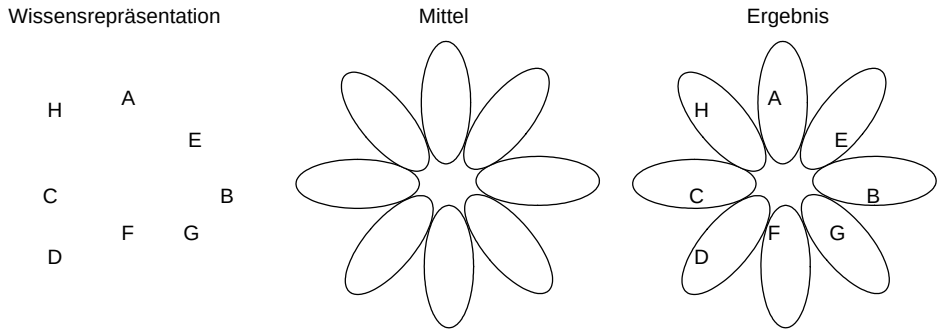
[265] Einteilung von Voß J., (2004), Begriffssysteme – Ein Vergleich verschiedener Arten von Begriffssystemen und Entwurf des integrierenden Thema-Datenmodells, Studienarbeit im Diplomstudiengang Informatik, HU Berlin, Version 1.1d, S. 26. http://www.nichtich.de/epub/begriffssysteme03/begriffssysteme.pdf :8.3.2004, S. 7. Auch wenn es sich formal nur um eine Studienarbeit handelt, enthält der Text doch eine interessante interdisziplinäre und vor allem aktuelle Bearbeitung des Themas.

[266] „Knowledge Representation is the area of AI (artificial intelligence) concerned with the formal symbolic languages used to represent the knowledge data used by intelligent systems and the data structures used to implement those formal languages. ... In most cases the intended use is to use explicitly stored knowledge to produce additional explicit knowledge. This is what reasoning is." Shapiro S.C., Artificial Intelligence, S. 54-57 in: Shapiro S. C. (Hrsg.) (1992), Encyclopedia of Artificial Intelligence, 2. Aufl. John Wiley & Sons Inc. New York 1992.

## 2. Überblick über Wissensorganisationssysteme

Man sollte zwischen dem Wissen selbst, seiner Repräsentation, dem Vorgang des Ordnens von Wissen (Systematisierung), dem Instrument oder Mittel der Wissensordnung (Begriffssystem, Datenmodell) und dem Ergebnis der Ordnung (Wissen in organisierter, formalisierter Form) unterscheiden (Grafik 36).

**Grafik 36: Ein Wissensraum, ein präkoordiniertes Klassifikationsschema (Mittel) und das klassifizierte Wissen (Ergebnis)**




Ohne diese Unterscheidung entsteht leicht Verwirrung. So kann man z.B. unter einem Cluster sowohl das (leere oder dokumentunabhängige) Mittel der Einteilung wie auch die (gefüllten) Haufen verstehen. Selbst der Vorgang bzw. die Vorgehensweise heißt ‚clustern'.

Dieses Kapitel konzentriert sich grundsätzlich auf das **Mittel der Organisation**, das aber nicht immer getrennt von den Inhalten existiert. Ein Begriffssystem kann dem Wissen implizit eingeschrieben werden, so dass die Wissensordnung überhaupt nur als Ergebnis greifbar ist.[267] Dies ist vor allem bei einfacheren Begriffsystemen der Fall, also etwa bei einer Termliste oder einem einfachen Glossar: Die Liste mit den Termen enthält implizit eine bestimmte Ordnung durch die Anordnung der Terme. Anders als im natürlichsprachlichen Kontext wird ein Term nach dem anderen angeordnet, z.B. nach dem Alphabet oder nach der Häufigkeit. Wenn ein Glossar nur einen oder gar keinen Term enthält, dann kann auch die Ordnungsstruktur nicht erkannt werden. Dazu müsste eine Art Dokumentationsregel zur Glossarerstellung vorhanden sein, die selbst aber kein Glossar mehr ist. Ein Cluster (Haufen) ohne Elemente hat keine Struktur, da ein Haufen noch gar nicht existiert. Auch hier kann man die Regeln des Clusterns beschreiben.

Eine Wissensorganisation heißt **präkoordiniert**, wenn das Mittel der Organisation entworfen wird, bevor man es auf das zu ordnende Wissen anwendet. In Grafik 2 ist das Ordnungsmittel so regelmäßig, weil es vorab entworfen wurde. Die Folge ist, dass die Daten nicht immer gut ins Schema passen. Andererseits kann dasselbe Schema öfter verwendet werden und Daten können in dieser Ordnung ausgetauscht werden. Ein typisches präkoordiniertes Wissensordnungssystem ist die Dewey Decimal Classification (DDC). Typisch **postkoordiniert** ist das Wissen in Wortfeldern organisiert, denn ein Wortfeld entsteht erst bei der Untersuchung der konkreten Wörter. Ein Cluster vereint beide Ansätze, denn das Clusterprogramm bekommt Kriterien vorgegeben, nach denen es die Wissensrepräsentanten untersuchen und zu Haufen zusammenfassen soll. Diese Kriterien können sehr allgemein bleiben (‚immer *next neighbour*') oder sehr detailliert werden (‚bilde acht nichthierarchische Haufen mit gleich vielen Elementen').

[267] Die Beispiele für Begriffssysteme von Voß a.a.O. sind daher geordnete Wissensräume, Produkte, und nicht Instrumente. Dort wird nach dem Inhalt unterschieden (Wörterbuch enthält eher Lemmata, Glossar eher Definitionen), nicht nach der Ordnungsstruktur. Eine inhaltliche Unterscheidung ist für die Untersuchung von Wissensordnungen nicht sinnvoll, weil die Grenzen fließend, subjektiv und für die Ordnung nicht ausschlaggebend sind.

Die vorgestellten Wissensorganisationssysteme (siehe Überblickstabelle 16 auf DIN A3) sind nach der **Komplexität** der verwendeten künstlichen Sprache in Vokabular und Syntax geordnet. Je komplexer Wissen symbolisiert werden kann, umso weiter unten steht das Wissensordnungssystem in der Tabelle. Die konkrete Realisierung desselben Wissensorganisationssystems kann einmal sehr einfach sein und ein andermal sehr komplex. Ein Index kann z.B. sehr einfach einige Wörter aus Dokumenten angeben, er kann aber als thematische Ordnung aller Dokumentabschnitte mit einem gebundenen Vokabular auch sehr komplex sein.

Die folgende Liste von Instrumenten zur Wissensorganisation ist weder abschließend noch ausführlich, aber sie macht höchst augenfällig, dass (rechtliches) Wissen auf viele verschiedene Weisen geordnet werden kann. Die Organisationssysteme kommen aus verschiedenen Wissenschaftszweigen, so dass dieser Überblick multidisziplinär und heterogen ist.

**Tabelle 16: Begriffsklärung von 21 Wissensorganisationssystemen**

| Was gibt es? | Weitere (englische) Bezeichnung | Definition oder Beschreibung | Zweck oder Beispiel |
|---|---|---|---|
| **Schlagwörter** | *keyword* | Eine Zeichenfolge aus dem Dokument, die es beschreibt | Themensuche in Datenbanken |
| **Lexikalische Ketten** | *lexical chains* | Lexikalische Ketten verbinden miteinander verwandte Textwörter im Themenbereich eines Textes wodurch sich die Wörter gegenseitig auf eine Bedeutung festlegen. | Textzusammenfassung, Disambiguierung. Herauszufinden mittels lexikalischer Datenbank, Prototypensemantik, Pronomenauflösung. Spiegeln Koreferenz, *Bridging* und Kontiguität. |
| **Normdatenbank** | Autoritätsdatei, *authority list*[A.] | Ergebnis der Normalisierung von Bezeichnungsvarianten zu kontrollierten Indexbegriffen | Eigennamen[B] zusammenführen. |
| **Kontrollierte Sprache** | *controlled language* | Sprache mit eingeschränktem Wortschatz und eingeschränkten Formulierungsregeln. | Erhöhte Eindeutigkeit der Sprache fördert Lesbarkeit und Übersetzbarkeit. Anwendung: Produktdokumentationen |
| **Denotative Begriffe**[C] | *linguistically-motivated indexing, descriptive term selection* | Verwalten nicht-hierarchischer, aber eindeutiger, meist normalisierter Namensräume. | Persönliche Informationsverwaltung. Beispiel für Namensräume: Insel Java, Programmiersprache Java oder "Java_Insel" und "Java_Programmiersprache". |
| **Strukturierte Schlagwörter** | Gebundene Schlagwörter, *structured keywords*. | Inhaltsbeschreibende Wörter plus deren Beziehung zueinander. | Digitale Büchereien |
| **Cluster** | *clustering* | Automatisches Zusammenfassen ähnlicher Dokumente zu Dokumenthaufen (Cluster). | Fachgebietseinteilung bei der Bistro-Korpussuche |
| **Inhaltsangabe** | *summary*[D]. Typen: Inhaltsverzeichnis, Auszug, Zusammenfassung am Ende, Kurzreferat (*abstract*)[E], Rezension. | Inhaltliche Kurzfassung von Text für die Zwecke eines Nutzers oder einer Anwendung. | AutoSummarize im Textverarbeitungsprogramm WORD (im Tools Menü) |
| **Index** | Verzeichnis, Register, Liste | Inhaltliche oder sachliche Indizierung von Dokumenten für eine gezielte Wiedergewinnung mit frei wählbarem oder gebundenem Vokabular. Das Vokabular kann z.B. alphabetisch oder thematisch sortiert sein. | Inhaltssuche in Printmedien |
| **Strukturierter Index** | *structured index, systematic index* [F], gebundener Index. | Zweckorientierte intellektuelle Einteilung von Dokumenten zu Rubriken | Dewey Decimal Classification (DDC), Universal Decimal Classification (UDC), Medical Subject headings (MeSH). |
| **Latent semantischer Index** | *latent semantic index*[G] | Automatische Einteilung von Zeichenketten zu Zeichenhaufen: (*Mutual Information*-) Clustering | (automatisches) *Information Retrieval*, Suchmaschinen Vivisimo[H] und Excite |
| **Taxonomie** | *taxonomy* | Netzwerk der Beziehungen „gehört zu“ oder „besteht aus“ zur Klassifizierung. | (Natur-, Sprach-, Finanz-,...) Wissenschaft |
| **Dendrogramm**[I] | Clusteranalyse in einem Baumdiagramm | Baumdarstellung der hierarchischen Zerlegung einer Datenmenge in immer kleinere Teilmengen. | Repräsentation hierarchischer Datenballen[J] (Cluster). Genotypanalyse bei der Züchtung. |
| **Wissenskarte** | Knowledge Maps, Kmaps | Visualisierung von Metawissen über Wissensquellen und Wissensträger. | Visualisierung von Wissen mit Hypermedien, Workflow Systeme, *Groupware*, Intranet. |
| **Themenkarte** | Topic Map | Zentrales Dokument einer Sammlung, das die Inhalte und Beziehungen aller anderen Dokumente anwendungsunabhängig beschreibt bzw. eine Dokumentsammlung, die um eine SGML Topic Map aufgebaut ist.[K] | Recherche: Techquila's Topic Map World Topic Map[L] |
| **Thesaurus**[M] | Synonymwörterbuch | Ein Thesaurus enthält kontrolliertes, natürlichsprachliches Indexvokabular, mit dem Konzepte oder Themen in formal eindeutige Beziehungen eingeordnet werden.[N] | Schnittstelle zwischen Benutzer und Information. (Automatisches) *Information Retrieval* in einem durch Benutzerkreis, Art der Dokumente und Erschließungstiefe begrenzten Bezugsrahmen. |
| **Termlisten, Glossare, Wörterbücher**[O] | *list of terms, term list, glossary, dictionary* | Wörterbücher können in mehreren Sprachen oder einer künstlichen Sprache sein. Ordnungstechnisch interessant sind sprachneutrale Dokumentrepräsentationen durch eine Interlingua. | Maschinelle Übersetzung |
| **Wortfeld** | Semantisches Feld, Bedeutungsfeld, *semantic field, lexical field*, sprachliches Feld, Verwendungsweisenfeld | Lexikalisches Paradigma, das durch die Aufteilung eines lexikalischen Inhaltskontinuums auf inhaltsunterschiedliche Wörter entsteht. Zusammenfassung sprachlicher Mittel nach semantischen Beziehungen zu Feldern. | Theorie der Wortbedeutung und des Bedeutungswandels, Theorie der Wortschatzarchitektur, Sprachlernhilfen. |
| **Begriffssystem** | Begriffsplan, Begriffsfeld. | Menge von Begriffen, zwischen denen Beziehungen bestehen oder hergestellt worden sind und die derart ein zusammenhängendes Ganzes darstellen.[P] | Terminologie, Lexikographie.<br>Beispiel BISTRO. |
| ***Concept Map***[Q] | *Conceptual map*,[R] Begriffslandkarten, Begriffsnetzdarstellungen, *Mind Maps, Web Maps, Cognitive Maps, Cognitive Structures, Sensitive Map, theme map, tree map, self organizing maps*. | Personalisierte graphisch-verbale Darstellung von Bedeutung und Zusammenhängen von Konzepten (strukturierter Wissensinhalte). | kreative Prozesse, Brainstorming. |
| **Ontologie** | ontology | Explizite formale Spezifikation einer gemeinsamen Konzeptualisierung[S] | Erleichtert Austausch, Teilen und Übersetzen von Wissen sowohl zwischen Mensch und Maschine wie zwischen verschiedenen Systemen. Künstliche Intelligenz. |

[A] http://www.ericfacility.net/extra/pub/ialsearch.cfm : 28.10.2003
[B] http://www.tessmann.it/seiten/opac/index_einfach_de.html : 28.10.2003
[C] Denotation eines Begriffs ist alles, was mit dem Begriff benannt werden kann. Konnotation ist die kommunikative Bedingung der Begriffsverwendung.
[D] Nach dem Oxford Handbook of Computational Linguistics 2003 höchstens 50 % des Ausgangstextes.
[E] Ein indikatives Abstract gibt an, wovon ein Dokument handelt. Ein informatives Abstrakt gibt darüber hinaus an, wie und mit welchem Ergebnis ein Sachverhalt behandelt wird.
[F] http://www.jura.uni-duesseldorf.de/rave/e/ravesyse.htm : 28.10.2003
[G] Bereits Deerwester S., Dumais S. T., Landauer T. K., Furnas G. W., Harshman R. A. (1990), Indexing by latent semantic analysis, S. 391-407 in: Journal of the Society for Information Science 41(6).
[H] http://vivisimo.com: 17.11.2003.
[I] http://www.cs.uni-bonn.de/~arbuckle/abis/semi-automatic_results.html : 28.10.2003
[J] http://www.dbs.informatik.uni-muenchen.de/~kailing/Fopras/singleLink.html : 28.10.2003
[K] ISO/IEC 13250:2002(E): “a) A set of information resources regarded by a topic map application as a bounded object set whose hub document is a topic map document conforming to the SGML architecture defined by this International Standard. b)…“
[L] http://www.techquila.com/topicmaps/tmworld/ : 28.10.2003
[M] Definition **Thesaurus:** “The vocabulary of a controlled indexing language, formally organized so that the a priori relationships between concepts (for example as “broader” and “narrower”) are made explicit” (ISO 2788, 1986:2). "A controlled set of terms selected from natural language and used to represent, in abstract form, the subjects of documents" (ISO 2788, 1986:2). Beispiel: http://www.lexisnexis.com/infopro/products/index/thesABnew.shtml: 17.11.2003, http://bigmac.phil.uni-sb.de/trex: 17.11.2003.
[N] ISO 2788, 1986:2 und DIN 1463.
[O] Einen guten Überblick über künstliche bzw. konstruierte Sprachen (interlinguas) bietet http://en.wikipedia.org/wiki/Artificial_language: 5.11.2003.
[P] DIN 2331:2
[Q] Definition **Conceptual Map**: “Concept mapping involves identifying concepts or ideas pertaining to a subject, and then describing the relationships that exist between these ideas in the form of a drawing sketch. The purpose of the map is to represent an individual's understanding of a body of knowledge and to illustrate the relationships among ideas that are meaningful to him or her.” (Sherratt and Schlabach 1990: 60-61) “Schematic device for representing a set of concept meanings embedded in a framework of propositions.” (Novak and Gowin, 1984) **Unterschied zwischen Conceptual Maps und Cognitive Maps:** “Conceptual maps are related to the model of the thesaurus, cognitive maps are related to the mental model of users” (Borgman 1986). “A concept map is a map constructed by experts, and represents many of the commonalties of their knowledge of the domain […] a cognitive map is a map constructed by a non-expert individual […] and has a lesser degree of disciplinary validity” (Mahler et al. 1991). http://www.uni-saarland.de/fak5/ezw/abteil/lehr/concept/concept-map : 28.10.2003
[R] http://www.december.com/web/text/images/cyberland.gif : 28.10.2003
[S] Anders das Oxford Handbook of Computational Linguistics 2003: Inventar der Objekte oder Prozesse und (meist hierarchische) Beschreibung ihrer Beziehungen zueinander.

2.1. Schlagwörter

Schlagwörter sind wohl die einfachste Art der Organisation von Wissen. Dokumente werden durch einzelne Zeichenfolgen dargestellt. Man findet das Dokument über die Zeichenfolgen wieder. Mit Schlagwörtern werden Suchmaschinen bedient und sie sind auch in Bibliotheken und Datenbanken fast allgegenwärtig. Die Benutzer haben sich längst auf die informationstechnische Beschränktheit von Schlagwörtern eingestellt, kombinieren Schlagwörter und komponieren Suchen. Schlagwort oder Ansetzungsform nennt man aber auch terminologisch kontrollierte[268] Bezeichnungen, die für einen Begriff aus dem Dokumenteninhalt verwendet wird. Ein Schlagwort kann auch aus mehreren Wörtern bestehen, z.B. Vor- und Familienname, Namen von Körperschaften, bibliographische Angaben, Zeitangaben u.ä.

2.2. Lexikalische Ketten

Lexikalische Ketten nennt man die Verbindung von Zeichenfolgen zu einem Sinngefüge. In den Beispielen: Blatt–schreiben; Herbst–Blatt–Baum; Blatt–Renovierung–Barock werden drei Bedeutungen von „Blatt“ durch einen prototypischen Kontext unterschieden. Gleichzeitig werden dadurch auch die anderen Wörter auf eine Bedeutung festgelegt.

Die lexikalischen Ketten können also Zweideutigkeiten ausräumen oder Texte zusammenfassen. Sie müssen aber nicht auf den menschlichen Benutzer zugeschnitten sein. Welche Zeichenfolgen zusammengekettet werden[269] und die Beziehung der Zeichenfolgen zueinander bleibt offen. Teilweise wird verlangt, dass die verketteten Wörter eine semantische Verwandtschaft aufweisen müssen, um mehrdeutige Lesarten zu disambiguieren.[270] Die lexikalische Kohäsion wird z.T. durch Kookkurenz im Text errechnet,[271] teilweise mit weiteren Heuristiken aufbereitet[272]. Ein neuerer Ansatz richtet die lexikalische Verkettung am menschlichen Nutzer aus. Mit Hilfe mnemotechnischer Erkenntnisse bildet man ‚assoziative lexikalische Ketten’.[273]

Wenn lexikalische Ketten präkoordiniert verwendet werden, dann kann nur der Bruchteil an Wörtern automatisch geordnet werden, der mit dem anderen Wort gemeinsam verwendet wird, alle anderen müssen intellektuell zugewiesen werden. Postkoordinierte lexikalische Ketten sind nicht unbedingt für den Menschen verständlich.

2.3. Normdatenbank

Eine Normdatenbank enthält für einige Ausdrücke eine Ersetzungsvorschrift, nach der sie durch andere, erlaubte Ausdrücke ersetzt werden können. Die bundesdeutsche Schriftsprache schreibt ei-

---

[268] ‚Terminologisch kontrolliert’ kann heißen, dass alle Wörter in der Grundform notiert werden; dann handelt es sich nach DIN 2342 um Lemmata, eine Sonderform des Schlagworts.

[269] Benutzte Verfahren sind z.B. der Abgleich mit einer lexikalischen Datenbank, Prototypensemantik, Pronomenauflösung, spiegeln, Koreferenzberechnung, Bridging und Kontiguität.

[270] Runte M., Beißwenger M., Storrer A., Modellierung eines Terminologienetzes für das automatische Linking auf der Grundlage von WordNet. In: Kunze C., Lemnitzer L., Wagner A. (Hrsg.), GermaNet Workshop Proceedings Tübingen 2003. http://www.sfs.uni-tuebingen.de/lsd/GermaNet-Workshop/TermNet.pdf: 14.11.2003.

[271] Barzilay, R.; Elhadad, M.; Using lexical chains for text summarization. In: Proceedings of the Intelligent Scalable Text Summarization Workshop (ISTS 1997), ACL, Madrid, 1997, http://citeseer.nj.nec.com/barzilay97using.html: 14.11.2003.

[272] Brunn, M.; Chali, Y.; Pichak, C. J., Text Summarization using lexical Chains. In: Workshop on Text Summarization, New Orleans, Louisiana, USA 2001, http://citeseer.nj.nec.com/brunn01text.html: 14.11.2003. Silber K.F., McCoy H.G., Efficient Text Summarization Using Lexical Chains, in: Proceedings of the 5th International Conference on Intelligent User Interfaces, New Orleans, Louisiana (USA) 2000, S. 252-255, http://web.media.mit.edu/~lieber/IUI/Silber/Silber.pdf: 14.11.2003. Zuletzt mit einer Kombination des Brown Korpus mit WordNet: Teich E., Fankhauser P. (2004), WordNet for Lexical Cohesion Analysis, S. 326-331 in: Sojka P., Pala K., Smrž P., Fellbaum C., Vossen P. (Hrsgg.) (2004), Proceedings of the Second International WordNet Conference (GWC 2004), Masaryk Universität Brno 2003.

[273] Diese Idee wird, so weit ersichtlich, in englischsprachiger und deutschsprachiger Literatur z.Z. nur unter http://www.elama.de/assoc : 23.9.2004 angeführt.

nige ausdrücke mit ‚ß', in der schweizer Schriftsprache wird jedes ‚ß' durch ‚ss' ausgedrückt. Wenn Wissen repräsentiert werden soll, dann sollte man sich solchen Schreibvarianten Rechnung tragen. In der Suche oder bei der Extraktion von Wörtern aus Dokumenten wird dann die Normdatenbank verwendet, damit z.B. ein Schweizer die bundesdeutschen Dokumente findet. Oft wird ein Ausdruck zur erlaubten Variante erhoben und die Varianten (Schreibvarianten, Kurzform/Langform, Mehrzahl, Synonyme, Dublettenabgleich usw.) werden durch diesen kontrollierten Indexbegriff ausgedrückt. Durch die Normalisierung in der Vorarbeit wird der Index selbst kürzer und eindeutiger.[274] Wenn auch der Text der Dokumente selbst normalisiert wird, handelt es sich um *controlled language*.

2.4. Kontrollierte Sprache

Eine kontrollierte Sprache ist selbst kein Wissensorganisationssystem, sondern die zweckorientierte Einschränkung einer natürlichen Sprache. Dabei nähert man sich einer künstlichen Sprache an, indem man die Flexibilität bei der natürlichen Sprachbildung bewusst unterdrückt, um eine höhere Eindeutigkeit (für Mensch und Maschine) zu erzeugen und die Voraussetzungen zur richtigen Interpretation ihres Inhalts zu senken.[275] Durch die Verwendung kontrollierter Sprache werden Lexikon und Strukturen einförmig, die Information ‚zwischen den Zeilen' und die Textintention kann schlecht wiedergegeben werden. Man hofft, dass ein standardisierter und vereinfachter Text für Nichtmuttersprachler oder für Maschinen leichter zu interpretieren sei.

Kontrollierte Sprache kann überall Anwendung finden, wo bisher natürliche Sprache verwendet wurde,[276] also auch bei der Wissensorganisation. Dann werden die natürlichen Ausdrücke zur Repräsentierung der Textinhalte (Schlagwörter usw.) in kontrollierter Sprache ausgedrückt, und evtl. auch die Texte selbst. Wenn bereits die Texte in kontrollierter Sprache verfasst sind, dann wurde die Normdatenbank sozusagen bereits in den Text integriert. Dann müssen allerdings beim gesamten Umgang mit den Texten die Normen beachtet werden, also beim Erstellen, Importieren, Verändern und Verwalten von Texten.

2.5. Denotative Begriffe

Auch denotative Begriffe sind selbst kein Wissensorganisationssystem, aber sie sollen das in der natürlichen Sprache enthaltene Wissen leichter zugänglich machen und werden in der Wissensorganisation verwendet.[277]

Denotative Begriffe bezeichnen nicht eine Klasse von Dingen, sondern geben nur eine bestimmte Menge tatsächlicher Instanzen an. Der Begriff „Mutter von L." bezieht sich auf eine bestimmte Person, während „ Tante von L." sich auf eine Klasse von Personen bezieht, die eine bestimmte Eigenschaft aufweisen. Die Verwendung der Denotation von Begriffen ist eine Möglichkeit, natürliche Sprache einzuschränken. Sie wird bei einfacheren, oft persönlichen Wissensorganisationssystemen eingesetzt, um Eindeutigkeit ohne Hierarchiebildung zu erreichen. Eine Klassenbezeichnung kann

---

[274] Einen Überblick gibt Junger U., Staatsbibliothek zu Berlin, http://www.museumsbund.de/fgdoku/dmbdoku_termine/dmbokt2002/beitraege/normdaten_junger.pdf: 14.11.2003. Die Schlagwortnorm-Datei und die Personennorm-Datei sind Normdatenbanken der Deutschen Bibliothek. Erstere enthält Ansetzungsformen von Schlagworten und Körperschaftsnamen, letztere Ansetzungsformen von Personennamen.

[275] Kittredge R. I. (2003), Sublanguages and controlled languages, S. 430-447 in: Mitkov R. (Hrsg.) (2003), The Oxford Handbook of Computational Linguistics, Oxford University Press 2003. Die kontrollierte Sprache beruhe auf der Annahme, dass "technical jargon (sublanguage) and irregular writing of [...] domain experts needs to be clarified (e.g. disambiguated), standardized and interpreted for [...] not native speaking domain experts", S. 441-444.

[276] Einen guten Überblick über die Anwendung und den Stand der Wissenschaft bieten die Controlled Language Applications Workshops 1996, 1998, 2000 und 2003. Von Anfang an war maschinelle Übersetzung ein Hauptanwendungsgebiet.

[277] Überblick von Arampatzis A. , van der Weide Th.P., Koster C.H.A., van Bommel P., in: An Evaluation of Linguistically-motivated Indexing Schemes, Proceedings of the BCSIRSG 2000, http://citeseer.nj.nec.com/arampatzis00evaluation.html: 17.11.2003.

durch einen Eigennamen oder eine laufende Nummer auf eine Instanz eingeengt werden, so z.B. „Tante_Franziska“, „Sohn_Nr.2“.

2.6. Strukturierte Schlagwörter

Die bisherigen Ordnungen waren alle gleichgeordnet. Strukturierte Schlagwörter sind inhaltsbeschreibende Ausdrücke in einer Hierarchie.

Ein gutes Beispiel für die Verwendung strukturierter Schlagwörter ist die Suche in der Datenbank des *Workplace Safety and Insurance Appeals Tribunal.*[278] Die Suche läuft in zwei Schritten ab. Zuerst sucht man mit einem Stichwort in den Schlagwörtern, man gibt z.B. „smok“ ein. Man erhält alle Schlagwörter, die diese Buchstabenfolge enthalten und sucht dann das passende aus der Liste aus. Im Beispiel wurde „*smoking*“, “*smoking (second-hand smoke)*”, “*respiratory condition*” und “*cancer (lung)*” angeboten. Die Schlagworte sind zum einen unterschiedlich spezifisch (*smoking* als Überbegriff), und den Ausdruck „*second-hand smoke*“ konnten auch englische Muttersprachler nicht unbedingt als den einzigen Ausdruck für Passivrauchen erwarten. Auf diese Weise wird mehr Wissen in die Wissensrepräsentation integriert.

2.7. Cluster

Cluster sind nicht ausschließlich ein Wissensorganisationsinstrument. Clustering ist die Teilung von Daten in mehrere, einander auf irgendeine Weise ähnliche Datenmengen. In einer grafischen Darstellung sehen die erstellten Mengen wie eine Traube (*cluster*) aus.[279] Man kann die Dokumente selbst oder die Ausdrücke der Wissensrepräsentation clustern. Man kann die Anzahl der Mengen auch vorher festlegen und man kann sogar jede Menge zunächst mit einem Idealbeispiel besetzen, nach dem dann die anderen Daten zugeordnet werden. Entscheidend für den Erfolg ist das verwendete Ähnlichkeitsmaß und ob sich die inhaltlichen Unterschiede in derselben Weise auch in der Darstellung wieder finden. Problematisch ist oft die Evaluierung, der Fehlernachweis und die Begründung der Einteilung, da sie auf Datenebene erfolgt und das Verhältnis der Daten zum Wissen meist völlig ungeklärt ist.[280] Cluster bieten sich an, wenn viele Datensätze einzuteilen sind und wenn jeder einzelne Datensatz sehr groß ist.

2.8. Inhaltsangabe

Eine Inhaltsangabe ist die inhaltliche Kurzfassung von Text für die Zwecke eines Nutzers oder einer Anwendung.[281] Das Wissen wird zwar in einer kompakteren Form dargestellt, aber weiterhin durch einen natürlichsprachlichen Text. Nach der Definition handelt es sich also nicht um eine Wissensrepräsentation im informationstechnischen Sinn (vergl. Fußn. 258). Das Wissen ist zwar kompakter, liegt aber weiterhin in natürlicher Sprache verschlossen. Die Ordnung des Wissens wird nicht anders symbolisiert, so dass menschliche Nutzer zwar eine Zeitersparnis haben können, aber für eine automatische Verarbeitung wenig gewonnen ist.

---

[278] Das Workplace Safety and Insurance Appeals Tribunal ist das letztinstanzliche Gericht für Arbeitsplatzsicherheit und arbeitsversicherungsrechtliche Streitigkeiten in Ontario (Kanada), http://www.wsiat.on.ca: 17.11.2003. Die Suche mit strukturierten Schlagwörtern ist unter http://www.wsiat.on.ca/ExtDec/KeywordDirectory.asp : 23.9.2004 zu finden.

[279] Cluster (englisch für: Haufen, Traube, Klumpen, Anhäufung, Gruppe) müssen in sich homogen und untereinander heterogen sein.

[280] In Kledowetz M., Symbolische Charakterisierung relationaler Datenbestände, Diplomarbeit Informatik TU Chemnitz, http://archiv.tu-chemnitz.de/pub/2002/0127: 4.11.2003 wird versucht, den unbenannten Datenhaufen wieder sinntragende Namen zuzuordnen und dadurch interpretierbar zu machen. Dazu wird den Dokumentgruppen symbolische Namen aus einer gegebenen Ontologie automatisch zuzuweisen.

[281] Siehe Überblick in Hovy E. H., Text Summarization, in: The Oxford Handbook of Computational Linguistics, Mitkov R. (2003)(Hrsg.), Oxford University Press 2003, S. 599-615.

Die intellektuelle Anfertigung von Inhaltsangaben ist sehr aufwendig und deshalb wird seit den 50er Jahren versucht, diese Aufgabe automatisch zu lösen. Auch hier ist die Evaluierung[282] schwierig, weil die Anforderungen sehr unterschiedlich sein können (Schwierigkeit des Inhalts; mehrere Texte zusammenfassen oder einen; mehrere Sprachen zusammenfassen; wie stark soll gekürzt werden; darf extrahiert werden oder soll neuer Text kreiert werden; Stichworte oder kohärenter Text). Außerdem hängt die Nützlichkeit einer Inhaltsangabe zur Wissensorganisation auch von den anderen Inhaltsangaben ab. Ein neu hinzukommendes, ähnliches Dokument kann eine gute Inhaltsangabe nachbesserungsbedürftig machen, weil nun auf die Unterschiede zum neu hinzugekommenen Dokument eingegangen werden muss.

Obwohl Inhaltsangaben für Menschen gemacht werden, könnte man sie in andere Wissensorganisationsmethoden einbinden. Die einfachste Kombination ist eine Volltextsuche in den Inhaltsangaben mit anschließender Zugriffsmöglichkeit auf das Dokument, falls das Informationsbedürfnis nicht schon gedeckt ist.[283]

2.9. Index

Ein Index[284] ist eine systematische intellektuelle Sammlung von Schlagworten oder Stichworten[285]. Die Stichworte können inhaltlich, sachlich, alphabetisch und gemischt geordnet werden. Sie können auf ein Dokument, auf mehrere oder auf Abschnitte eines Dokuments verweisen. Verschiedene Aspekte der Indizierung können in einem oder in getrennten Indizes verwirklicht werden. Dem entsprechend ist die Komplexität des integrierten Wissens in einem Index sehr verschieden. Ein Schlagwortindex spielt bei Printmedien die Rolle einer verkürzten Volltextsuche.

In der Informationswissenschaft werden die Stichwörter auch als Deskriptoren bezeichnet und können auch in einer künstlichen Sprache sein. Für die Vergabe von Deskriptoren (das Indexieren)[286], gibt es die DIN 31623 über die „Indexierung zur inhaltlichen Erschließung von Dokumenten“[287].

2.10. Strukturierter Index

Ein strukturierter Index liegt vor, wenn die systematische Schlagwortsammlung eine weitere Ordnung aufweist. Dazu kann eine Hierarchie eingeführt werden, z.B. Rubriken (*directories*).[288] In manchen Bereichen konnte sich eine Indexierungsstruktur so stark durchsetzen, dass sie wie ein

---

[282] Siehe die Text Summarization Evaluation Conference SUMMAC, http://www.itl.nist.gov/iaui/894.02/related_projects/tipster_summac: 17.11.2003 und Inderjeet M., Summarization Evaluation: An Overview, http://citeseer.nj.nec.com/591246.html: 17.11.2003.

[283] Ein Beispiel für eine Rechtsdatenbank, die Inhaltsangaben verwendet, ist die Rechtsprechungsdatenbank des deutschen Städtetags, http://extranet.staedtetag.de/suche/index.html: 17.11.2003. Man kann dort in Entscheidungstext, Pressemitteilung, Inhaltsangabe oder Leitsatz getrennt suchen. Auch das Auswärtige Amt organisiert das in den Auslandsvertretungen erarbeitete Wissen in einer Kaskade von inhaltszusammenfassenden Berichten.

[284] Es gibt den Ausdruck Index auch für die Bezeichnung eines Belegstellenwörterbuchs, das Lemmata und deren Belegstelle in einer Textgrundlage enthält. Bekannt ist das Goethe-Wörterbuch (GWb), ein textbezogenes Bedeutungswörterbuch, das Goethes aktiv schriftlich verwendeten Wortschatz vollständig erfassen wird, http://bibliothek.bbaw.de/Goethe/my_html/wortschatz.htm: 17.11.2003.

[285] Während ein Schlagwort eine terminologisch kontrollierte, evtl. mehrwortige Bezeichnung ist, sind in einem Stichwortindex Wörter des Dokuments so aufgeführt, wie sie im Text erscheinen.

[286] Einen kompakten und praxisorientierten Überblick bietet Umstätter W., Nutzen der Indexierung bei Online-Datenbanken, 14. Online- Tagung 1992 Frankfurt am Main, DGD-Schrift (OLBG-13) 2/92 S. 403-420, http://www.ib.hu-berlin.de/~wumsta/pub65.html: 17.11.2003. Umstätter betont, dass Indizes die Wissensvernetzung erleichterten.

[287] DIN 31623, Indexierung zur inhaltlichen Erschließung von Dokumenten, Berlin 11/78 insbes. Teil 3: Syntaktische Indexierung mit Deskriptoren (Sept. 88, 4 S.), Beuth Verlag Berlin 1988.

[288] Bei Printmedien ist das Branchenbuch das klassische Beispiel, in der EDV die Webverzeichnisse von Suchmaschinen. Bei http://www.yahoo.com: 18.11.2003 werden Webseiten nach Themen, Aktualität, Nachrichten, Bildern und meistbesuchten Seiten sortiert.

Standard funktioniert.[289] Bekannte Beispiele sind die Dewey Decimal Classification (DDC) und die Universal Decimal Classification (UDC) für die Einteilung von Printmedien, sowie die Medical Subject Headings (MeSH) in der Medizin.[290] Im Internet werden solche strukturierte Indizes *information gateways* genannt.

2.11. Latent semantischer Index

Ein latent semantischer Index ist ein Index, der durch *(Mutual Information) Clustering* automatisch aufgebaut wird und von dem angenommen wird, dass seine Rubriken inhaltlich zusammenhängen. Der entscheidende Unterschied zu den bisher vorgestellten Indizes ist, dass der latent semantische Index nicht manuell erstellt wird. Der Unterschied zum Cluster ist, dass das Ziel nicht die Bildung von Gruppen ist, sondern eines hierarchischen Indexes, der möglichst nah an die Qualität eines manuell erstellten Indexes heranreichen soll.[291]

2.12. Taxonomie

Die Taxonomie[292] im engeren Sinn ist der Zweig der Systematik, der sich mit der Einordnung der Lebewesen in Taxa befasst. Eine Taxonomie im weiteren Sinn ist eine Menge von Begriffen, die durch Überbegriff-Unterbegriff-Beziehungen[293] miteinander verbunden sind.[294] Ein klassisches Beispiel ist die Klassifikation in der Biologie, wie sie in Tabelle 17 gezeigt wird.

**Tabelle 17: Klassifikation von Pflanzen - Bezeichnungen und Hierarchie von Kategorien und Taxa**

| Kategorie | | | Taxon (Beispiele) |
|---|---|---|---|
| deutsch | lateinisch | Endung | |
| Reich | regio | -ae | Plantae |
| Abteilung oder Stamm | divisio | -phyta | Spermatophyta Unterabteilung: Angiospermae oder Magnoliophytinae[295] |
| Klasse | classis | -phyceae (Algen) -atae / -opsida (Gefäßpflanzen) | Monokotyledonae (= Liliopsida) |
| Unterklasse | subclassis | -phycidae (Algen) -ideae (Gefäßpflanzen) | Commelinidae |
| Ordnung | ordo | -ales | Poales |
| Familie | familia | -aceae | Poaceae |
| Gattung | genus | | Poa |
| Art | species | | Poa annua L |

Tabelle zum einjährigen Rispengras nach P. v. Sengbusch.[296]

Ein hervorragendes Beispiel einer neueren Anwendung der Taxonomie ist die eXtensible Business Reporting Language (XBRL)[297] zur Vereinheitlichung von Finanzinformationen. Die XBRL

[289] Übersichten findet man unter http://www.public.iastate.edu/~CYBERSTACKS/CTW.htm: 17.11.2003, http://www.lub.lu.se/desire/sbigs.html: 17.11.2003. Ein bekannter Gateway ist http://bubl.ac.uk: 17.11.2003.

[290] Eine Informationssuche über mehrsprachige medizinische Inhaltsangaben wurde im Projekt „Multilingual Concept Hierarchies for Medical Information Organization and Retrieval (MuchMore)“ am Deutschen Forschungszentrum für Künstliche Intelligenz (DFKI) entwickelt: http://muchmore.dfki.de/ : 23.9.2004.

[291] http://javelina.cet.middlebury.edu/lsa/out/tutorial.htm : 28.10.2003.

[292] Zu griechisch *táxis* für Ordnung und *nomos* für Gesetz.

[293] Hyperonym – Hyponym oder „is_a“ Beziehungen.

[294] Vossen P., Ontologies, in The Oxford Handbook of Computational Linguistics, Mitkov R. (Hrsg.), Oxford University Press 2003, S. 464-482, S. 467.

[295] Herkömmliche Bezeichnungen werden beibehalten. Es gibt oft Untergruppierungen wie subclassis, subordo, subfamilia, subgenera und subspesies (ssp.). In vielen Gattungen werden Sektionen (sectiones) und in vielen Familien Triben (tribus) unterschieden.

[296] http://www.biologie.uni-hamburg.de/b-online/e43/t1.htm : 2.2.2004.

definiert Taxa von Finanzinformationen (z.B. die Positionen von Bilanzen) und ihre Beziehungen zueinander (zum Beispiel, dass das Umlaufvermögen eine Position der Aktiva ist). Damit wird ein (erweiterbarer) nationaler und internationaler Sprachstandard geschaffen, der dank der gleichzeitigen Definition in einer elektronischen Sprache den Kommunikationsfluss beschleunigt, qualitativ verbessert und kostengünstiger macht.[298]

2.13. Dendrogramm

Ein Dendrogramm ist die grafische Darstellung einer hierarchischen Systematik, bei der auf jeder Hierarchieebene die relative Ähnlichkeit bezüglich eines weiteren Kriteriums durch die Entfernung der Taxa (Knoten) voneinander angegeben wird. Entsprechende dreidimensionale Darstellungen heißen Kegelbaum (*Cone Tree*, früher C*am Tree*). Das Dendrogramm dient in erster Linie der Visualisierung von Information und steht hier stellvertretend für eine ganze Reihe von Verfahren.[299]

Wenn nicht jeder Knoten mit Objekten besetzt ist, muss aus den Objekten aller untergeordneten Knoten ein virtueller Ort des Knotens bestimmt werden, Zentroid genannt.[300] Geeigneter für eine Darstellung im Dendrogramm sind daher Einteilungen, bei denen jeder Knoten mit Objekten besetzt ist.

2.14. Wissenskarte

Wissenskarten[301] visualisieren das Wissen darüber, wer wo Wissen hat. In einer landkartenartigen Übersicht sind aktuelle Wissenserzeuger, Wissensträger und Wissensaufbewahrungsstellen in strukturierter Weise verzeichnet. Es gibt derzeit keine Standards für eine solche Karte, die auch in einer Serie ineinander verschachtelter Karten verschiedenen Maßstabs bestehen kann. Zum Aufbau einer solchen Wissenskarte für eine Organisation geht man regelmäßig von den formalen Zuständigkeiten und Funktionen aus und bringt in Erfahrung, welches Wissen tatsächlich dort verortet ist.[302] Das Ergebnis weicht oft von der Theorie ab, wenn bestimmte Probleme noch nicht aufgetaucht sind, ein Personalwechsel stattgefunden hat oder das notwendige Wissen spontan andernorts beschafft wird. Dann wird dem an diesem anderen Ort vorhandenen Wissen nachgegangen und es wird in der Wissenskarte verzeichnet. Eine Wissenskarte kann Risiken aufdecken, wenn betriebswichtiges Wissen nur in einer Datenbank vorhanden ist oder sensibles Wissen weit gestreut ist. Wissenskarten zeigen, wo Wissensteile effizienter vernetzt, Wissensgenerierung gefördert und Personalentwicklung unterstützt werden kann und zeigt ganz allgemein an, wo das eigentliche Wissenskapital zu finden ist.

---

[297] Das X in XBRL verweist auf seine Notation in XML (eXtended Markup Language). XBRL ist sozusagen das Wörterbuch einer (konstruierten) Fachsprache der internationalen Finanzwissenschaft. Schmidt G., „XBRL ist... das, was die Finanzwelt braucht!“, in: Consultant - Steuern Wirtschaft Finanzen, 05/2002, Max Schimmel Verlag Würzburg, 2002.

[298] http://www.xbrl-deutschland.de: 19.11.2003. Die Deutsche Taxonomie GermanAP Version 1.0 liegt in http://www.xbrl-deutschland.de/GermanAP_2002_02_15_nav.htm: 19.11.2003.

[299] Ein gute Übersicht zur Informationsvisualisierung bietet Englberger H., Computergestützte Informationsvisualisierung - Eine Klassifikation aktueller Techniken und ihre Einsatzpotentiale für die Unternehmung, Diplomarbeit TU München 1995, http://www11.informatik.tu-muenchen.de/publications/pdf/da-englberger1995.pdf: 19.11.2003. Dort wird die Information Landscape als Technik primär zur Informationspräsentation besprochen, während primär zum Informationsretrieval folgende Techniken angeführt werden: Cone Tree, Perspective Wall, Stretch Tools, Information Grid, LyberWorld, InfoCrystal, Fractal Tree, Document Lens, Information Cube, Hyperbolic Tree, Spiral Calendar; und als Techniken primär zur Informationsexploration Tree Map, Bead, Data Sphere, VisDB, Table Lens, Data Visualization Sliders, HyperMap, IVEE und DataSpace.

[300] Deutlich wird dies in der grafischen Darstellung in Arbuckle T., http://www.cs.uni-bonn.de/~arbuckle/abis/semi-automatic_results.html: 19.11.2003.

[301] Zu Wissenskarten als Visualisierung von Metawissen über Wissensquellen und Wissensträger vergl. Nohr H., Automatische Indexierung, Einführung in betriebliche Verfahren, System und Anwendungen, Materialien zur Information und Dokumentation, Verlag für Berlin-Brandenburg, Potsdam, 2001, sowie ders. in http://www.iuk.hdm-stuttgart.de/nohr/Km/KmPubl/wisska/wisska_1.html: 19.11.2003.

[302] http://www.vtt.fi/ele/research/soh/projects/totem2001/ kmpmhtml/kmpm4.1.2.html : 3.2.2004.

2.15. Themenkarte

Themenkarten (*topic maps*)[303] bezwecken Aufzeichnung und Austausch von Informationen darüber, welche Struktur Informationen zu einem bestimmten Thema haben und welche Beziehungen zwischen verschiedenen Themen bestehen. Dazu stellt ISO/IEC 13250: 2000[304] eine standardisierte Notation zur Verfügung. Das Ergebnis, also mehrere mit dieser Standardnotation verbundene Dokumente, sind eine Themenkarte. Üblicherweise enthalten Themenkarten thematische Information sowie Information über die Beziehung von Themen. Die thematischen Informationen können jeglicher Art sein und sind regelmäßig in Form von Dokumenten vorhanden.

Die Kapitel dieser Arbeit könnten eigene Dokumente sein, die Informationen zu jeweils einem Thema enthalten, also etwa Wissensorganisation, Termextraktion oder Textindizierung. Darüber hinaus würde eine Themenkarte anzeigen, wie die Beziehung dieser Themen zueinander ist, also dass in dieser Arbeit die Termextraktion zur Textindizierung benutzt wird bzw. dass sich die Textindizierung auf vorherige Termextraktion stützt.

2.16. Thesaurus

„Ein Thesaurus im Bereich der Information und Dokumentation ist eine geordnete Zusammenstellung von Begriffen und ihren (vorwiegend natürlichsprachlichen) Bezeichnungen, die in einem Dokumentationsgebiet zum Indexieren, Speichern und Wiederauffinden dient. Er ist durch folgende Merkmale gekennzeichnet: a) Begriffe und Bezeichnungen werden eindeutig aufeinander bezogen („terminologische Kontrolle“), indem Synonyme möglichst vollständig erfasst werden, Homonyme und Polyseme besonders gekennzeichnet werden, für jeden Begriff eine Bezeichnung (Vorzugsbenennung, Begriffsnummer oder Notation) festgelegt wird, die den Begriff eindeutig vertritt, b) Beziehungen zwischen Begriffen (repräsentiert durch ihre Bezeichnungen) werden dargestellt.“[305]

Ein Thesaurus soll also das Instrument für den Benutzer sein, um zu Informationen zu gelangen. Je mächtiger dieses Instrument wird, umso komplizierter wird in der Regel seine Benutzung sein.[306]

2.17. Termlisten, Glossare oder Wörterbücher

Termlisten können zur Textindexierung eingesetzt werden (vergl. Kapitel 3). Fachgebiete und Glossare strukturieren Terminologie. Wörterbücher im weiteren Sinn geben die (Übersetzungs-)Beziehung zwischen Dokumenten wieder, die somit geordnet werden. Diese selten zum Zweck des Informationsmanagements erstellten Instrumente können durch Planung weiterentwickelt werden. Plansprachen können entweder Grammatik und Wortschatz natürlicher Sprachen übernehmen (aposteriorische Plansprache) oder sie werden neutral gegenüber natürlichen Sprachen konzipiert, so dass sie einen eigenen Wortschatz und ein eigene Grammatik haben (apriorische Plansprache.[307] Bereits Leibniz versuchte sich an der Konstruktion einer logischen Sprache. Termlisten, Glossare, Wörterbücher und dergleichen eignen sich vor allem zur Organisation von Wissen über die Sprache, zur Erzeugung künstlicher Intelligenz sowie für mehrsprachige Inhalte.[308]

---

[303] Freie Editorsoftware unter http://www.ontopia.net/download/freedownload.html : 12.3.2004.

[304] http://www.y12.doe.gov/sgml/sc34/document/0129.pdf : 2.2.2004.

[305] DIN 1463. Nach der ISO 2788, 1986:2 ist ein Thesaurus “the vocabulary of a controlled indexing language, formally organized so that the a priori relationships between concepts (for example as "broader" and "narrower") are made explicit" bzw. "a controlled set of terms selected from natural language and used to represent, in abstract form, the subjects of documents".

[306] Kostenlose und kommerzielle Software zur Thesauruserstellung: http://www.willpower.demon.co.uk/thessoft.htm : 11.3.2004.

[307] Plansprachen (englisch: *planned language, constructed language* oder *conlang*). Glück H., Schlagwort ‚Plansprache’, S. 532-533 in: Glück H. (2000), Metzler Lexikon Sprache, 2. Auflage J. B. Metzler Verlag Stuttgart Weimar 2000.

[308] Kommerzielle Terminologiesoftware unter http://www.iim.fh-koeln.de/dtp/werkzeuge.html : 11.3.2004, Shareware unter http://www.dicomaker.netfirms.com/ : 11.3.2004, freie Software und kostenlose Demoversionen unter http://www.lai.com/lai/tg.html : 11.3.2004.

2.18. Wortfeld

Ein Wortfeld ist ein lexikalisches Paradigma, das durch die Aufteilung eines lexikalischen Inhaltskontinuums auf inhaltsunterschiedliche Wörter entsteht.[309] Es werden also alle sprachlichen Mittel, die zum Ausdruck von Bedeutung verwendet werden können, in einem zusammengehörigen Feld dargestellt. Die Darstellung erfolgt meist grafisch oder in Tabellenform[310].

Das Wortfeld veranschaulicht sprachliches Tiefenwissen zur Unterscheidung sinnverwandter Wörter (z.B. verschiedener Sprachen) und ist damit zur Einordnung sinnverschiedener Wörter in ein einziges System weniger geeignet. Wortfelder sind auf ihren Bedeutungsbereich begrenzt und nicht miteinander vernetzt.

2.19. Begriffssystem

Ein Begriffssystem ist eine Menge von Begriffen, zwischen denen Beziehungen bestehen oder hergestellt worden sind und die derart ein zusammenhängendes Ganzes darstellen.[311] Im Gegensatz zum Wortfeld liegt der Akzent auf der zusammenhängenden und erschöpfenden Darstellung in der Breite, während die Begriffe selbst als gewusst vorausgesetzt werden. Das System steht hier im Vordergrund. Im Gegensatz zur Taxonomie ist ein Begriffssystem tendenziell eher postkoordiniert und datenorientiert.

2.20. *Concept Map*

Eine *Concept Map* ist eine personalisierte grafisch-verbale Darstellung von Bedeutungen und Zusammenhängen von strukturierten Wissensinhalten. Die Zusammenhänge sind also subjektiv und stellen die Vorstellungen eines Individuums dar. *Concept Maps* setzen mehrere Ausgangspunkte zueinander in Beziehung, während *Mind Maps* nur einen zentralen Ausgangspunkt haben. [312]

Die *Mapping*-Technik eignet sich besonders zum Erlernen, Erinnern und Entwerfen.[313] Die Individualität persönlicher *Maps* gerät beim Austausch mit anderen unter Konformitätsdruck und nähert sich dann einem objektiven Begriffssystem an.[314] Der Zweck von *Maps* ist nicht die Ordnung objektiven, vorhandenen rechtlichen Wissens.

2.21. Ontologie

Eine Ontologie im informationswissenschaftlichen Sinn[315] ist die explizite formale Spezifikation einer gemeinsamen Konzeptualisierung.[316] Das Ziel einer Ontologie ist die Wiedergabe eines objektiven oder zumindest intersubjektiven Ausschnitts der Welt durch eine formale Sprache, die zumin-

---

[309] Für Wortfeld werden auch die Bezeichnungen Bedeutungsfeld, semantisches Feld, Sinnbezirk, Begriffsfeld, Sachfeld, sprachliches Feld, lexikalisches Feld und Verwendungsweisenfeld benutzt. Das Wortfeld ist „eine Menge von partiell synonymen Wörtern bzw. Lexemen, d.h. Lexemen mit einem gleichen bzw. ähnlichen Inhalt bzw. Bedeutungskern." Schaeder B., Schlagwort ‚Wortfeld', S. 798 in: Glück H. (2000), a.a.O.

[310] Solch eine Tabelle gibt die Bedeutungsmerkmale von sprachlichen Mitteln an, um sie zu differenzieren.

[311] ISO 1087 - Terminology / Vocabulary (1990), "3.10 system of concepts: Structured set of concepts established according to the relations between them, each concept being determined by its position in this set." Siehe auch DIN 2338 Teil 1 (Juli 1984) und DIN 2331 - Begriffssysteme und ihre Darstellung (1980).

[312] „Beim *Mind Mapping* werden ausgehend von einem Thema als Mittelpunkt weitere Einzelheiten und Ideen als Abzweigungen notiert, von denen wiederum weitere Zweige und Unterverzweigungen abgehen." Schwarze, M. (2002), http://www.learn-line.nrw.de/angebote/selma/foyer/projekte/hammproj4/difference.htm : 3.2.2004. Eine informative Seiten zu diesem neuen und teilweise verwirrend umfangreichen Gebiet ist http://www.denkzeichnen.de/: 3.2.2004.

[313] Freie und kommerzielle Software zur Erstellung von *Mind Maps* u.ä.: http://www.mindmap.ch/software.htm : 11.3.2004.

[314] Der speziell auf das Erlernen rechtlicher Inhalte zugeschnittene Versuch in Hansen, M., Möglichkeiten des Einsatzes von *Concept-Maps* zum juristischen Lernen, JurPC Web-Dok. 103/1998 hat keinerlei individuelle Züge mehr, sondern ist gestalterisch wie inhaltlich ein Begriffssystem.

[315] Zu den verschiedenen Definitionen sehr lehrreich ist http://beat.doebe.li/bibliothek/w00085.html : 3.2.2004.

[316] Gruber, T. R. (1993), Towards Principles for the Design of Ontologies Used for Knowledge Sharing, in Guarino N., Poli, R., (Hrsg.), Formal Ontology in Conceptual Analysis and Knowledge Representation, Kluwer Academic Publishers Deventer (NL) 1993, zit. nach S. 1 in http://citeseer.nj.nec.com/gruber93toward.html : 21.9.2004.

dest Konzepte und Beziehungen präzise beschreibt. Die künstliche Sprache ermöglicht maschinelle Kommunikation, irrtumsfreie Übertragbarkeit und Generierung von Wissen (künstliche Intelligenz, siehe oben bei 2.17.) Durch eine Ontologie kann auch die Wissenskommunikation Mensch-Mensch und/oder Mensch-Maschine erleichtert werden.

### 3. Einteilung von Wissensorganisationssystemen für Recht

Welche Eigenschaften können Wissensorganisationssysteme haben, wie kann man die beschriebene Vielfalt vergleichend einteilen und wie findet man das nützlichste System für rechtliches Wissen?

Zunächst ist hervorzuheben, dass Wissensorganisationssysteme, die Rechtliches enthalten, auch alle **rechtliches Wissen** repräsentieren. Rechtswissen als Inhalt/Konzept/Thema der Wissenseinheiten ist z.B.: „Wer einen Anspruch hat, kann von einem anderen ein Tun oder Unterlassen verlangen." Das Rechtswissen gibt hier einen Tatbestand (Anspruch) und eine Rechtsfolge.

Eine andere Art von Wissen ist das **terminologische Wissen**, das darüber informiert, wie ein Konzept lexikalisiert wird, z.B.: „Anspruch kann man mit *pretesa* übersetzen. Ein Anspruch ist das Recht, von einem anderen ein Tun oder Unterlassen zu verlangen." Das Wissen hat eine lexikalische Information zum Inhalt. Auf die Ordnung solchen Wissens wurde bereits im Kapitel zur Fachgebietsklassifikation eingegangen. Die Wissensordnung von Rechtsbegriffen muss nicht und wird in der Praxis auch nicht mit der Ordnung rechtlichen Wissens übereinstimmen.

Häufig trifft man Wissensordnungen an, die nur mittelbar das rechtliche Wissen ordnen, indem sie unmittelbar **Dokumente ordnen**. Wenn in Datenbanken formale Merkmale des gespeicherten Wissens wie Dateiformat, Größe, Erstellungsdatum, Häufigkeit der Verwendung oder Bearbeitungsstatus verzeichnet sind, dann werden damit Dokumente geordnet, das Wissen selbst bleibt aber in den Dokumenten verschlossen. Bsp.: „Urteil Az. ... des Gerichts XY vom YZ, Titel: ‚Anspruch auf Unterlassung'." Eine Recherche in so geordneten Dokumenten eröffnet effektiven Zugriff auf das Wissen.[317]

Welche Art von Wissen geordnet werden soll, hat Folgen für die Wahl der Ordnungssytematik. Rechtliches Wissen hat bereits objektiv vorgegebene Strukturen, z.B. steht die Verfassung über normalen Gesetzen. Um diese Eigenschaft rechtlichen Wissens wiederzugeben, empfiehlt sich eine präkoordinierte Wissensordnung (zur Präkoordination s.o. S. 258).

Terminologisches Wissen hingegen richtet sich nach dem wissenschaftlichen Ansatz: Wenn bei einem monodirektional rechtsvergleichenden Ansatz Termini des deutschen Rechtssystems den Termini des Italienischen zugeordnet werden, dann wird diese gewählte Struktur die spezifische Ordnungsweise bestimmen (Postkoordination). Die Folge ist, dass monodirektional deutsch-italienische Daten nicht mit monodirektional italienisch-deutschen Daten in derselben Ordnungsstruktur verwaltet werden können.[318]

Soweit rechtliches Wissen also universal und objektiv ist oder von anderen jederzeit anerkannt[319] wird (Verfassung vor normalen Gesetzen), bietet sich diese Struktur zur Ordnung an. Bereits die Zuordnung von Strafrecht unter das öffentliche Recht oder als eigenständiges Rechtsgebiet ist eine subjektive Entscheidung und sollte keinen Einfluss auf die Organisation der Daten haben. Man sollte daher ein gewisses Problembewusstsein für die zeitliche und räumliche Relativität rechtlichen Wissens mitbringen. Selbst weit verbreitete Klassifikationen wie die DCC sind Revisionen unter-

[317] Vergl. hierzu das Kapitel zur Dokumentklassifikation.

[318] Selbst manche kommerzielle Terminologiesoftware setzt Übersetzungsbeziehungen absolut, so dass sie nicht auf eine Richtung beschränkt werden können. Inkompatible wissenschaftliche Ansätze werden in eine gemeinsame Ordnung gepresst.

[319] Ein gutes Beispiel für eine rein konventionelle aber allgemein anerkannte Einteilung ist die ICD10, eine internationale Klassifikation von Krankheiten. Vergl. http://www.med-kolleg.de/icd/index.htm : 12.3.2004.

worfen, bei denen sich die Klassen ändern und die dazugehörigen Dokumente neu sortiert werden müssen, was mit großem Aufwand verbunden ist. Nur die Revision erlaubt die Weiterentwicklung von Sachgebieten. Veränderliche Bezugsrahmen oder Aufgabenstellungen sollten daher von Anfang an in die Überlegungen zur Wissensorganisation einbezogen werden.

Ein weiteres Kriterium ist, ob eine **Hierarchie** des Wissens entsteht wie bei strukturierten Schlagwörtern, einem strukturierten Index, einer Taxonomie, einem Dendrogramm, einem Begriffssystem und einer Ontologie, oder ob das Wissen ohne Hilfe von Polen wie oben - unten, Zentrum - Peripherie geordnet wird wie bei einfachen Schlagwörtern oder bei einem Cluster. Es gibt auch verschiedene Formen der Hierarchie wie die strenge, linkseindeutige Monohierarchie, die Polyhierarchie, Heterarchie/zyklische Hierarchie[320], assoziative Ordnung und die sequentielle Hierarchie. Im Kapitel 2 zur Termdarstellung wurde bereits gezeigt, dass es nicht unbedingt nötig ist, Daten hierarchisch zu speichern, um eine Hierarchie anzeigen zu können. Die Hierarchie kann zur Darstellung auch dynamisch aus Datenrelationen erzeugt werden. Diese Lösung bietet sich vor allem dann an, wenn nicht nach einem einzigen Kriterium geordnet wird wie nach der genetischen Ähnlichkeit bei einem Dendrogramm. So eine eindimensionale Wissensordnung ist für rechtliches Wissen nicht geeignet.[321]

Ein wichtiger Punkt ist auch, inwieweit eine Wissensordnung kontrolliertes Vokabular und künstliche Syntax verwendet. Als Richtschnur kann man sagen: je künstlicher, umso stärker ist die Kommunikation auf Maschinen ausgerichtet statt auf Menschen. Künstliche Syntax kann alle mathematischen Relationseigenschaften darstellen. Während Enzyklopädien allgemein auf weiterführende Informationen verweisen und ihre Assoziationsrelationen[322] nicht weiter definieren, enthält ein Thesaurus zusätzlich fest definierte Äquivalenz- und Hierarchierelationen. Auch diese Relationen können aber präziser formuliert werden, denn Äquivalenz kann exakte Äquivalenz bedeuten, aber auch auf einen äquivalenten aber weiteren oder engeren Begriff verweisen.[323] Präzise kann auch die Gerichtetheit der Beziehung definiert werden (Äquivalenz nur in eine Richtung wie beim monodirektionalen Ansatz oder in beide Richtungen?). Schließlich sind sogar Relationen über Relationen mathematisch möglich.[324] In diesen Fragen ist stets der Aufwand bei der Erstellung der Relationen und der erwartete Nutzen durch logische Operationen und Interoperabilität abzuwägen. Nachdem noch kein Relationenformalismus standardisiert wurde, fällt vor allem der projektinterne Nutzen ins Gewicht.

Nachdem hier einige Wissensorganisationssysteme nach ihrer Geeignetheit für komplexe Zusammenhänge vorgestellt wurden und darauf hingewiesen wurde, dass ihre konkrete Verwendung davon oft erheblich abweicht, wird nun eine typische Rechtsdatenbank vorgestellt, für die ein Ord-

---

[320] Die meisten Informationen in unserer Umgebung sind hierarchisch organisiert (Stammbaum, traditionelle Printmedien, Computerverzeichnisse) und damit nicht zyklisch. Hypertext enthält oft Ringverweise und bildet dann ein zyklisches Netz.

[321] Weitere Kriterien je nach Ausgangsdaten können sein: Gibt es Darstellungsprobleme bei Überschneidungen? Soll die Nähe von Subklassen aussagekräftig sein? Gibt es viele verschiedene Arten von Klassen? Gibt es Standards für die Art von Wissen? Erlangt man das Wissen bereits mit einer bestimmten Ordnung versehen (Vergleiche Kapitel über die Dokumentklassifikation)? Wie soll das Wissen verwendet werden und könnte sich diese Anwendung ändern? Ist Aufwand oder Nutzen größer, wenn man das Wissen mehrfach strukturiert?

[322] Assoziationsrelationen sind nach DIN 2331, 3: Ähnlichkeitsrelation, Nachfolgerelation (Vorgänger - Nachfolger), Kausalrelation (Ursache - Wirkung), genetische Relation (Produzent - Produkt), Herstellungsrelation (Material - Produkt), Transmissionsrelation (Sender - Empfänger), instrumentelle Relation (Werkzeug - Anwendung des Werkzeuges), funktionelle Relation (Argument - Funktion), Antonymie-Relation (Gegensätze).

[323] Interessante Ergebnisse zur Äquivalenz bietet das MACS-Projekt, in dem die drei großen Bibliothekskataloge Schlagwortnormdatei (SWD), *Library of Congress Subject Headings* (LCSH) und *Répertoire d'autorité-matière encyclopédique et alphabétique unifié* (RAMEAU) aufeinander abgebildet wurden (*mapping*). http://infolab.uvt.nl/prj/macs/macsd.html : 15.3.2004.

[324] Hierzu und zum ordnungstechnischen Aspekt von Relationen aufschlussreich Voß a.a.O. S. 26-32.

nungssystem gesucht und implementiert wurde. Anhand dieses Fallbeispiels können Schwierigkeiten und Lösungsansätze in der Praxis am besten erläutert werden.

**4. Wissensorganisation in einer Rechtsdatenbank**

Rechtliches Wissen ist typischerweise in Rechtstexten enthalten (vergl. Einleitung). Viele Rechtsdatenbanken enthalten nur Rechtstexte eines bestimmten Typs, z.B. nur Gesetze oder nur Urteile. Die Texte sind oft alle in derselben Sprache und aus demselben Rechtssystem, z.B. deutsche Texte bundesdeutschen Rechts. Viele Rechtsdatenbanken spezialisieren sich noch weiter, wie die umfangreiche Gesetzsammlung des Bundesministeriums der Justiz[325] auf Bundesgesetze, oder die Datenbanken deutscher Gerichte[326] auf eigene Entscheidungen. Die Ordnungskriterien sind dort durchweg formaler Art (Datum, Aktenzeichen, Name des Gesetzes) oder minimal inhaltlich wie die Einteilung der Bundesgesetze in die Kategorien Rechtspflege, Verfassungs- und Verwaltungsrecht, Zivilrecht, Strafrecht, Handels- und Wirtschaftsrecht oder die Einteilung der Entscheidungen des Landgerichts Kassel[327] in Zivilsachen, Mietsachen, Beschwerden (Zivil), Kostenrechtsprechung, Strafsachen und Strafvollstreckung. Daneben steht fast immer eine Volltextsuche zur Verfügung.

Diese Auswahl zeigt, dass der gezielte Einsatz von Wissensorganisationssystemen im Recht noch kaum verbreitet ist. Dabei ist das Recht ein ideales Anwendungsgebiet für komplexe Wissensorganisationssysteme, weil das mit Texten und Fachbegriffen beschriebene Gebiet keine naturwissenschaftliche Wirklichkeit ist wie z.B. Pflanzenarten oder Krankheiten, sondern z.T. die soziale Wirklichkeit, vor allem aber eine Deontologie beschreibt, also einen geforderten sozialen Zustand. Die korrekte Beschreibung einer Bedeutung kann also nicht einfach durch Vergleich mit der Wirklichkeit verifiziert werden, sondern erfordert stets erneut eine rechtliche Argumentation. Damit befindet man sich bei der Beschreibung von Recht in einer vergleichbaren Situation wie beim Sprechen über Sprache. Die Rechtssprache ist aufgrund ihrer deontologischen Ausrichtung fast immer von der Ontologie und damit dem konkreten Gebrauch getrennt. Das Wort ‚Vertrag' wird kaum je für ein bestimmtes Vertragsdokument verwendet wie in der Geschichtswissenschaft. Meist wird der Begriff benutzt, um menschliches Verhalten generell Einzuordnen um das geforderte Verhalten daraus abzuleiten. Die Rechtssprache ist also regelmäßig nicht Objektsprache, sondern Metasprache, womit die Beschreibung der Rechtssprache meist Metasprache höherer Ordnung ist.[328] Diese Ausgangslage spricht für den Einsatz komplexerer Wissensinformationssysteme.

Ebenso deutet die Palette wünschenswerter Anwendungen elektronischer Speicherung von rechtlichem Wissen klar auf komplexe Beschreibungsmethoden hin. Rechtliches Wissen wird nicht nur über mehrere Sprachen gespeichert und gesucht, es wird auch in besonderem Maße übersetzt, bildet Ausgangspunkt für logische Argumentationen und soll die bedeutendsten Entscheidungen unterstützen.

---

[325] http://bundesrecht.juris.de : 5.2.2004.

[326] Eine Liste aller Gerichte mit Internetauftritt findet sich unter http://www.jura.uni-sb.de/internet/gericht.html : 5.2.2004. Einige typische Datenbanken mit Entscheidungen: Bundesverfassungsgericht http://www.bundesverfassungsgericht.de/cgi-bin/link.pl?entscheidungen : 5.2.2004, Bundesgerichtshof http://www.bundesgerichtshof.de/entscheidungen/entscheidungen.php : 5.2.2004, Bundesverwaltungsgericht http://www.bverwg.de/enid/926d526ba0f8166e40d164301a1aa626,0/96.html : 5.2.2004, US-American Legislative Information on the Internet of the Library of Congress http://thomas.loc.gov/ : 9.3.2004.

[327] http://www.lgkassel.de/ : 5.2.2004.

[328] In der englischen Sprachphilosophie heißt die konkrete Ebene *use* (Gebrauch) und die Metaebene *mention* (Erwähnung).

Auch das Projekt MIRIS (Minority Rights Information System)[329] trägt diese Kennzeichen der Komplexität des Rechts. MIRIS dokumentiert im Zuge des gewachsenen Bewusstseins und der zunehmend rechtlichen Behandlung von zumeist ethnischen, nationalen oder sprachlichen Minderheiten deren rechtliche Situation. Minderheitenfragen gewinnen mit der Rechtsvereinheitlichung in Europa, der Erweiterung der EU und der zunehmenden Austragung von Minderheitenkonflikten auf der internationalen Bühne[330] an Bedeutung. Dieses stark wachsende Rechtsgebiet mit internationalen, nationalen und nichtstaatlichen Akteuren macht eine systematische Zusammenstellung aller relevanten Rechtsdokumente wünschenswert. Dabei soll die rechtliche Situation von allen Minderheiten der Mitgliedstaaten des Europarats in Gesetz und Rechtssprechung aufgezeigt werden.

Die Zielgruppe von MIRIS sind alle, die vom Minderheitenrecht betroffen oder mit ihm befasst sind, also die Minderheiten selbst, staatliche Stellen, nichtstaatliche Minderheitenorganisationen und Wissenschaftler. Die Daten sind frei über das Internet zugänglich.

Andere Datenbanken decken entweder nur einige Länder ab (MINELRES[331] und CEDIME-SE[332] etwa Südosteuropa) oder bestimmte Themen (Mercator[333] nur Sprachminderheiten). Auch in diesen Datenbanken sind die rechtlichen Dokumente vorwiegend ungeordnet abgelegt. MIRIS enthält vor allem staatliche Gesetze (Grafik 37) im Hypertextformat (Grafik 38):

[329] http://www.eurac.edu/miris : 24.2.2004. Zu technischen Aspekten siehe Dongilli P., Gamper J. (2003), MIRIS Unleashed, S. 131-139 in: Proceedings JURIX 2003: The 16th Annual Conference on Legal Knowledge and Information Systems December 2003, IOS Press Utrecht 2003.

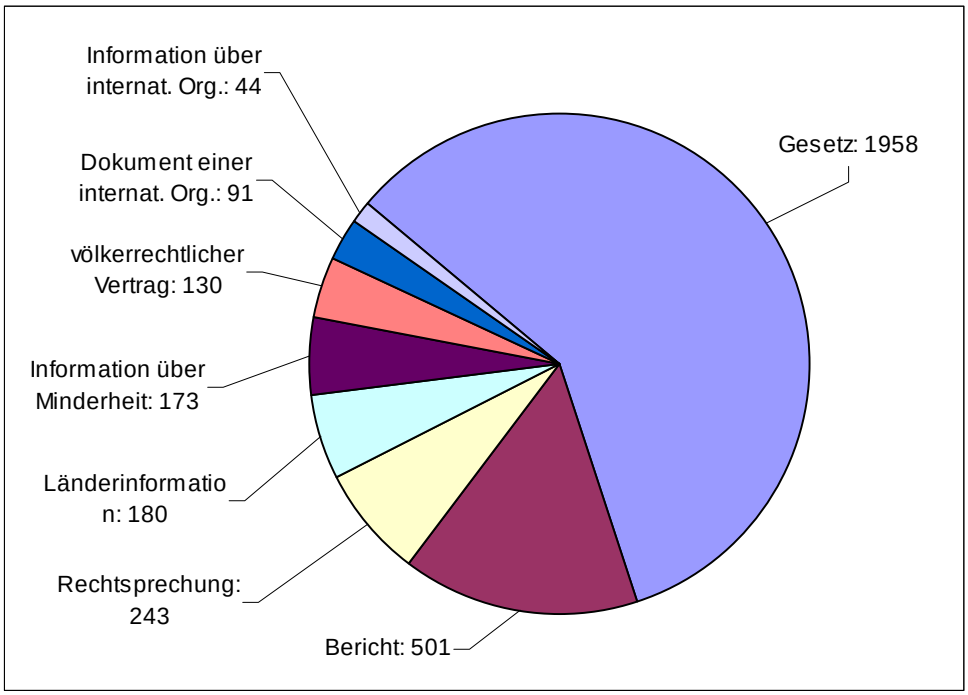
[330] Z.B. das Rahmenübereinkommen zum Schutz nationaler Minderheiten.

[331] Directory of resources on minority human rights and related problems of the transition period in Eastern and Central Europe: http://www.minelres.lv/ : 24.9.2004.

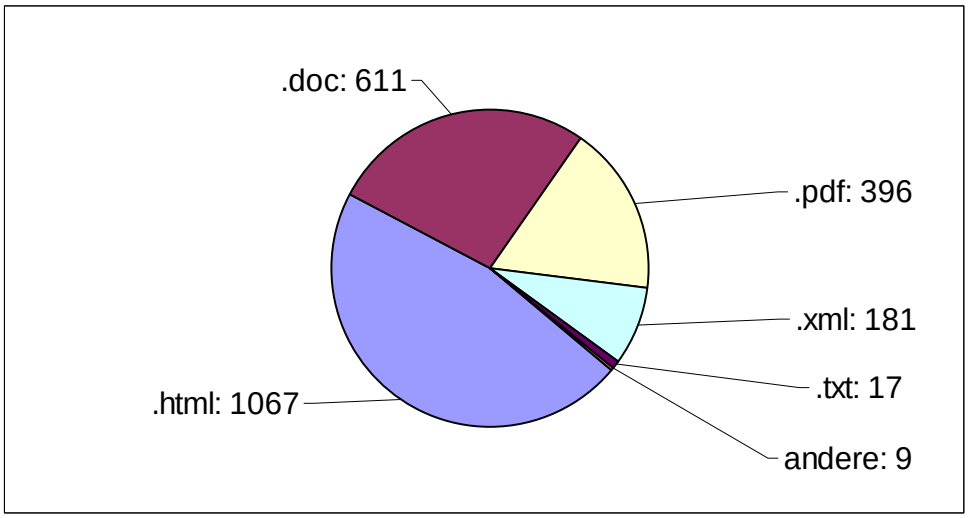
[332] Center for Documentation and Information on Minorities in Europe - Southeast Europe (CEDIME-SE): http://www.greekhelsinki.gr/bhr/english/index.html : 24.9.2004.

[333] Mercator ist ein Netz von drei Forschungs- und Dokumentationszentren mit dem Auftrag, der Bevölkerung mehrsprachiger Regionen zuverlässige objektive Informationen über die Minderheitensprachen zu bieten. Mercator-Education der Fryske Akademy (Ljouwert) untersucht die Stellung von Minderheitensprachen an Schulen; Mercator-Media der Aberystwyth Universität Wales in den Medien und Mercator-Legislation der CIEMEN Stiftung (Barcelona) dokumentiert die Sprachgesetzgebung und den offiziellen Sprachgebrauch: http://www.ciemen.org/mercator/index-gb.htm : 24.2.2004.

**Grafik 37: Formale Einteilung der Dokumente in MIRIS**



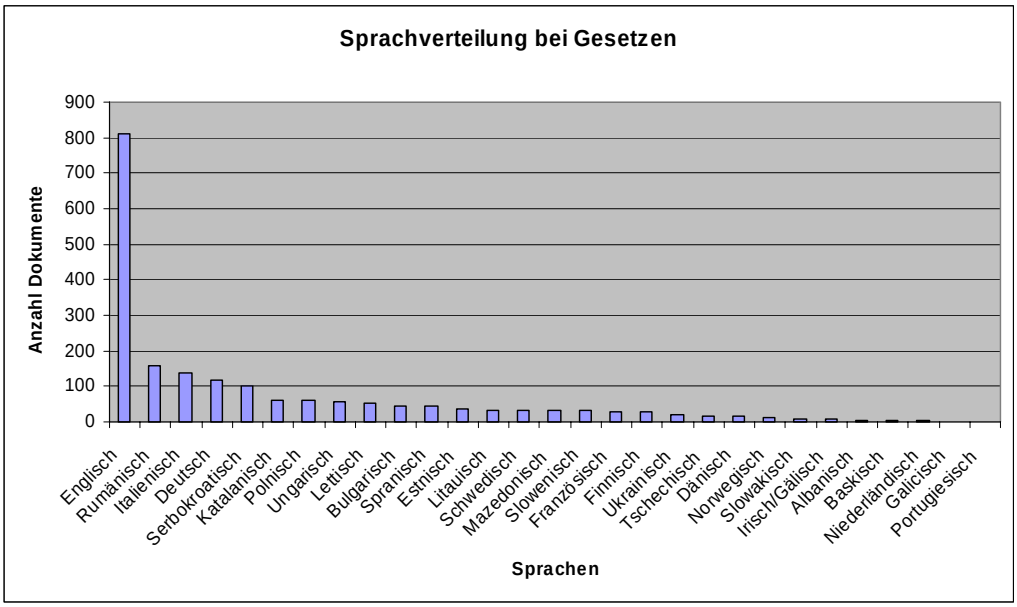

**Grafik 38: Dateiendungen der Dokumente in MIRIS**




Nicht alle Dokumente sind lokal gespeichert, weil sich Rechtstexte oft ändern, so dass ein Verweis zu dem externen Dokument effizienter ist. Außerdem entstehen in manchen Fällen Copyrightprobleme. Nur die Rechtssprechung ist hauptsächlich lokal gespeichert, weil sie selten dauerhaft unter einer Adresse gespeichert bleibt.

Interessant ist die Verteilung der Sprachen über die Dokumente (Grafik 39):

**Grafik 39: Sprachverteilung bei Gesetzen:**




Staatliche Gesetzgebung und Rechtssprechung entstehen natürlich in der offiziellen Staatssprache, zur Rechtsvergleichung werden sie aber oft übersetzt. Völkerrechtliche Verträge benutzen meist von vornherein mehrere große Sprachen. Staaten stellen oft Übersetzungen ihrer Verfassung in eine Weltsprache oder von minderheitenrelevanten Texten in die Minderheitensprache zur Verfügung. In Ausnahmefällen werden Rechtstexte auch speziell für MIRIS übersetzt.[334] Daher sind zwar viele ‚kleine' Sprachen wie Galicisch und Katalanisch vertreten, die Mehrheit der Texte ist aber auf Englisch abgefasst.

Zunächst wurden die Daten natürlich über eine Volltextsuche und eine Suche über die formalen Metadaten aufgeschlossen. Damit werden Dokumente zuverlässig gefunden, wenn ihre Existenz bekannt ist. Diffusere bzw. inhaltlich orientierte Informationsbedürfnisse („alles zum Thema XY") kann nur über die sprachgebundene Volltextsuche bedient werden. Vollständigkeit und Präzision lassen bei dieser Suche zu wünschen übrig, denn um alle richtigen Ergebnisse zu finden, müssten alle möglichen Verbalisierungen des Themas in allen gewünschten Sprachen gesucht werden, also z.B. Sprachminderheit, sprachliche Minderheit, *linguistic minority*, *language minority*, usw. Die so erhaltenen Treffer enthalten dann viele falsche Treffer und Dokumentübersetzungen.

Die Tatsache, dass die Rechtsdatenbank MIRIS Dokumente in vielen Sprachen enthält, verbessert die Volltextsuche, wenn alle gesuchten Treffer aus derselben Sprache sind, weil Treffer in den anderen Sprachen praktisch ausgeschlossen sind. Andererseits muss für jede Sprache extra gesucht werden.

Abgesehen von der Mehrsprachigkeit ist MIRIS typisch für Rechtsdatenbanken. Es gibt zu viele, zu verschiedene und zu häufig wechselnde Dokumente, als dass der Benutzer die relevanten Dokumente individuell ansteuern könnte. Der Nutzen der Datenbank hängt damit entscheidend von den explizit inhaltlichen Suchmöglichkeiten über das Wissensorganisationssystem ab.

[334] Etwa das kroatische Verfassungsgesetz über das Recht nationaler Minderheiten in der Republik Kroatien, Gesetzesblatt Kroatiens 155/2002.

### 5. Eine Ontologie für MIRIS

Wie oben dargelegt kann die Datenbank MIRIS durch eine Wissensbeschreibung verbessert werden, die komplexe Themen korrekt darstellt und dadurch die Vollständigkeit und Präzision von Dokumentsuchen erhöht.

Die Suche in der Wissensordnung sollte dem Benutzer die Themen und ihre Beziehung (zumindest Oberbegriff, Unterbegriff, verwandter Begriff) zueinander anzeigen. Die grafische Darstellung einer solchen Hierarchie ist aus Gründen der Übersichtlichkeit fast unabdingbar, und die Steuerung sollte intuitiv und einfach sein.

Wünschenswert wäre darüber hinaus, dass die Suche nicht mit der Auswahl eines Themas endet, sondern dass die Trefferliste bearbeitet werden kann. Das Informationsbedürfnis des Nutzers kann auf mehrere Themen bezogen sein oder auf bestimmte Dokumente zu einem Thema. Daher sollte es möglich sein, mehrere Themen zusammen auszuwählen, die Suche innerhalb eines Themas mit den Metadaten oder der Volltextsuche verfeinern, oder bestimmte Dokumente (etwa bereits konsultierte Dokumente) von der Liste zu streichen.

Schließlich sollte die Themenbeschreibung einerseits für alle Benutzer intuitiv und leicht verständlich sein, sich andererseits aber so weit wie möglich an Standards halten, damit der enorme Aufwand bei der Dokumentklassifizierung zugleich Vorteile beim Datentransfer bringt.

Insgesamt sollte der Arbeitsaufwand so gering wie möglich gehalten werden, wobei insbesondere zu beachten ist, dass die Technik zur Erstellung, Bedienung und Wartung des Wissensorganisationssystems einfach gehalten wird und so weit wie möglich von den Juristen des Projekts MIRIS selbst geleistet werden kann.

### 6. Ontologieerstellung: Knoten und Beziehungen

Diese Bedürfnisse sollte im Projekt MIRIS eine Ontologie befriedigen. In Zusammenarbeit mit der Freien Universität Bozen wurde zunächst das Gerüst der Themen oder Konzepte erstellt, denen dann in einem zweiten Schritt die Dokumente zugeordnet werden sollten.

Als Standard für die Themen- bzw. Ressourcenbeschreibung wurde das Metadatenformat Dublin Core (*Dublin Core metadata element set* DCMS) gewählt, dessen Version 1.1 u.a. dem ISO-Standard 15836:2003(E) entspricht.[335] Damit sind aber nur 15 Kategorien vorgegeben. Komplette logikbasierte Sprachen für Ontologien sind RDF/RDFS, DAML und OIL (kombiniert zu OWL) und der Topic Maps Standard.[336]

Damit die menschlichen Benutzer zumindest das Vokabular intuitiv verstehen, wurde als Klassenbezeichnungen eine natürliche Sprache gewählt. Für MIRIS bietet sich Englisch an, weil sich das Thema Minderheiten an ein internationales Publikum wendet und weil die thematische Suche die vielen englischsprachigen Dokumente ordnen soll. Abgesehen von den fünf großen Sprachen können die 50 oder weniger Dokumente in den seltener verwendeten Sprachen zur Not auch einzeln durchgesehen werden. Sie können auch als Originaltexte zu ihren englischen Übersetzungen gefunden werden.

Die Beschreibungssprache ist also Englisch. Beschrieben wird aber das Konzept, es handelt sich also um einen onomasiologischen Ansatz. Hier wird die Beziehung zwischen Rechtsdatenbanken und Rechtsterminologiedatenbank deutlich: Beide können unabhängig voneinander existieren, für eine gründliche Beschreibung des rechtlichen Wissens sind sie aber aufeinander angewiesen. Die

---

[335] "The Dublin Core metadata element set is a standard for cross-domain information resource description." Dublin Core Metadata Initiative, http://dublincore.org/documents/dces/ : 24.2.2004. Dublin Core 1.1 entspricht den Standards ISO 15836-2003 (Februar 2003): http://www.niso.org/international/SC4/n515.pdf ; NISO Standard Z39.85-2001 (September 2001): http://www.niso.org/standards/resources/Z39-85.pdf ; Europäisches Komitee für Normung, (EKN oder CEN) Workshop Agreement CWA 13874 (März 2000): http://www.cenorm.be: 2.3.2004.

[336] Zu diesen und weiteren Standards siehe Voß a.a.O., S. 36-43.

inhaltliche Beschreibung von Rechtsdokumenten benötigt eine Terminologie und die terminologische Beschreibung von Rechtskonzepten geschieht regelmäßig durch konkrete Verwendungsbeispiele und oft durch Rechtstexte in einem Korpus.

Die englische Terminologie zur Beschreibung der rechtlichen Ressourcen wurde aus drei Quellen kompiliert. Man zog existierende Rechtsthesauri heran, den lexikalischen Referenz- und Synonymthesaurus WordNet[337] sowie Fachbegriffe, die in der Wissenschaft verwendet und für das konkrete Projekt besonders geeignet erschienen.

Die sprachliche Beschreibung der Konzepte bezieht sich immer auf das zugrunde liegende Recht. Daher musste die Arbeit von Rechtsexperten gemacht werden. Außerdem muss die Ontologie an Rechtsänderungen und an die sich wandelnde Beschreibungstiefe und -breite in der Dokumentkollektion angepasst werden. Das Thema ‚Autonomie' mag für 50 Dokumente genügen, sollte bei 500 Dokumenten aber in ‚territoriale Autonomie', ‚finanzielle Autonomie' und ‚kulturelle Autonomie' unterteilt werden. Ebenso scheint das Thema ‚Homosexuelle' zunächst nicht zu den klassischen (nationalen, ethnischen oder sprachlichen) Minderheiten zu gehören, soweit deren Rechte aber wie die von klassischen Minderheiten und im selben Zusammenhang geregelt werden, können sie durchaus ein eigenes Minderheitenthema bilden. Ähnlich ist die Lage bei den sog. ‚neuen Minderheiten', die durch Migration entstehen, oder bei weiteren Gruppen, die sich des Minderheitendiskurses bedienen.

Als Technik zur Erstellung der Ontologie durch Juristen wurde das frei erhältliche Softwareprogramm Protégé-2000 verwendet.[338] Es erlaubt dem Benutzer, eine thematische Ontologie mit einer grafischen Oberfläche übersichtlich zu erstellen und zu bearbeiten, also die einzelnen Knoten und ihre Beziehungen zu definieren, ohne dafür Programmierkenntnisse haben zu müssen. Protégé-2000 stellt einerseits fertige Eingabemodelle bereit, andererseits transformiert es die Angaben unter Beachtung der jeweiligen Syntax in eine der gewählten Ausgabesprachen, z.B. in eine *Portable Network Graphic* (.png). Die Ontologie kann dann in jedem Browser als Hierarchiebaum angezeigt werden.[339] Noy et. al. beschreiben die technischen Möglichkeiten von Protégé.[340] Mit Dublin Core und Protégé wurde etwa das medizinische *semantic web* erstellt.[341] Die Grafiken 40 und 41 zeigen Abbildungen der grafischen Oberfläche:

[337] „WordNet® is an online lexical reference system whose design is inspired by current psycholinguistic theories of human lexical memory. English nouns, verbs, adjectives and adverbs are organized into synonym sets, each representing one underlying lexical concept. Different relations link the synonym sets. WordNet was developed by the Cognitive Science Laboratory at Princeton University under the direction of Professor George A. Miller (Principal Investigator)." http://www.cogsci.princeton.edu/~wn/ : 25.2.2004. WordNet 2.2 liegt unter der Adresse http://www.cogsci.princeton.edu/cgi-bin/webwn : 25.2.2004.

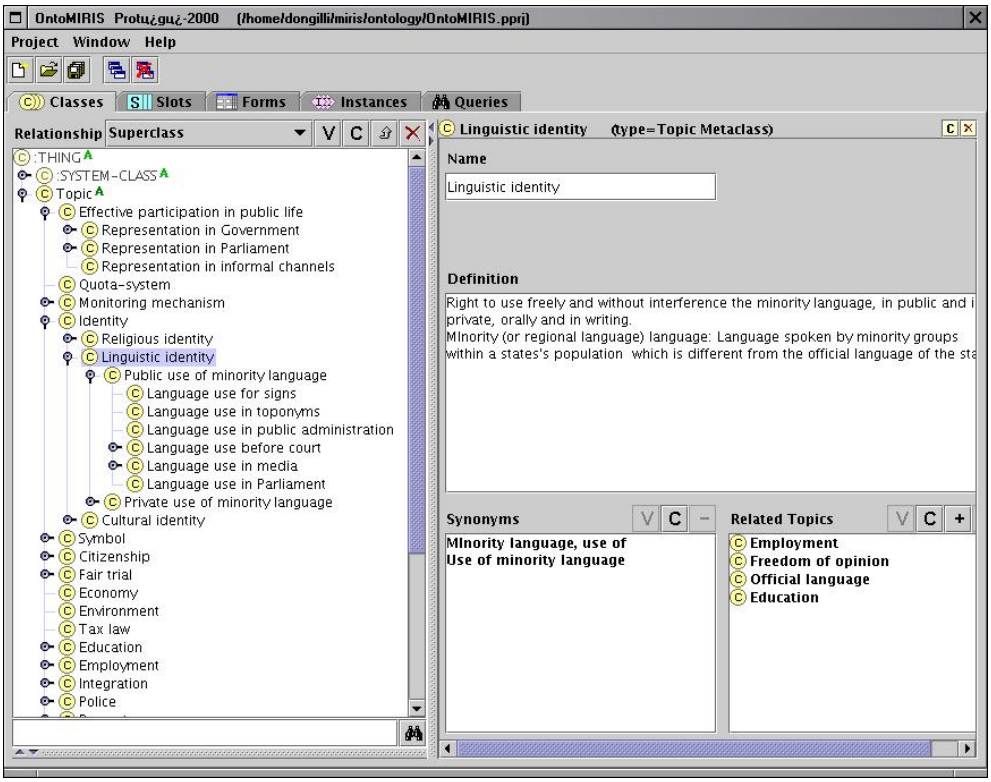
[338] http://protege.stanford.edu/ : 24.2.2004.

[339] http://dev.eurac.edu/~muser/ontomiris/ontoMIRIS-new.png : 2.3.2004.

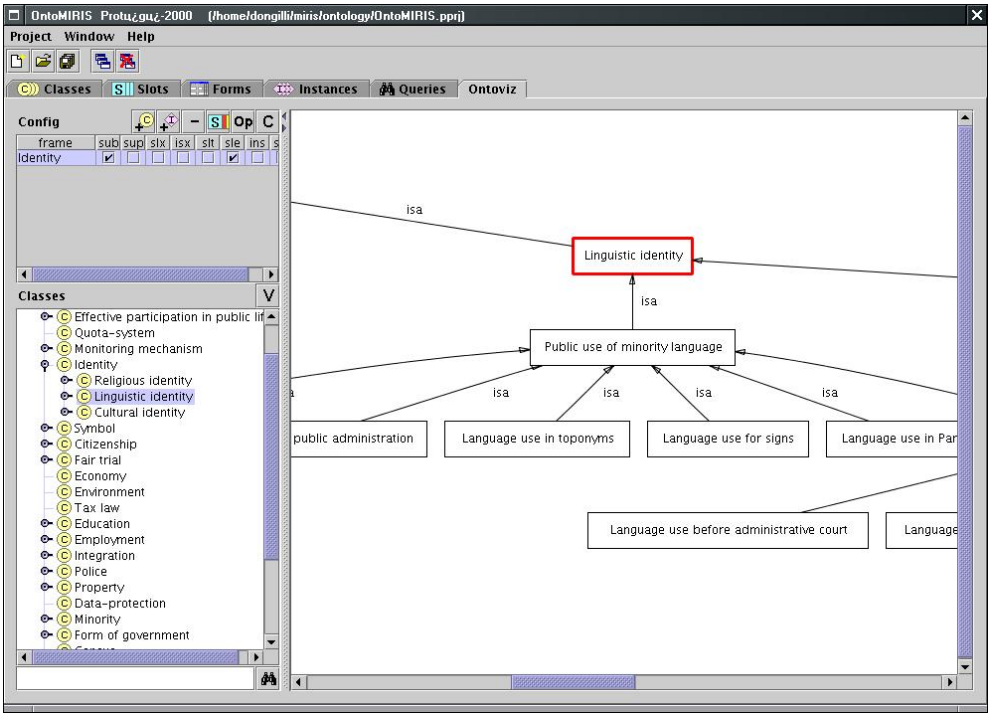
[340] N. F. Noy, M. Sintek, S. Decker, M. Crubezy, R. W. Fergerson, & M. A. Musen. Creating Semantic Web Contents with Protege-2000. IEEE Intelligent Systems 16(2):60-71, 2001. http://www-smi.stanford.edu/pubs/SMI_Reports/SMI-2001-0872.pdf

[341] http://www.touchgraph.com bietet interessantes Anschauungsmaterial, wie eine Ontologie graphisch in einem Browser dargestellt werden kann, etwa im PubMedBrowser V1.01 http://www.touchgraph.com/TGPubMedBrowser.html : 25.2.2004. Eine Zusammenfassung der Links zum semantic web gibt Sung-Jung Cho in http://ai.kaist.ac.kr/~sjcho/semantic-web/ : 2.3.2004.

**Grafik 40: Erstellung der Ontologieknoten mit Protégé-2000**

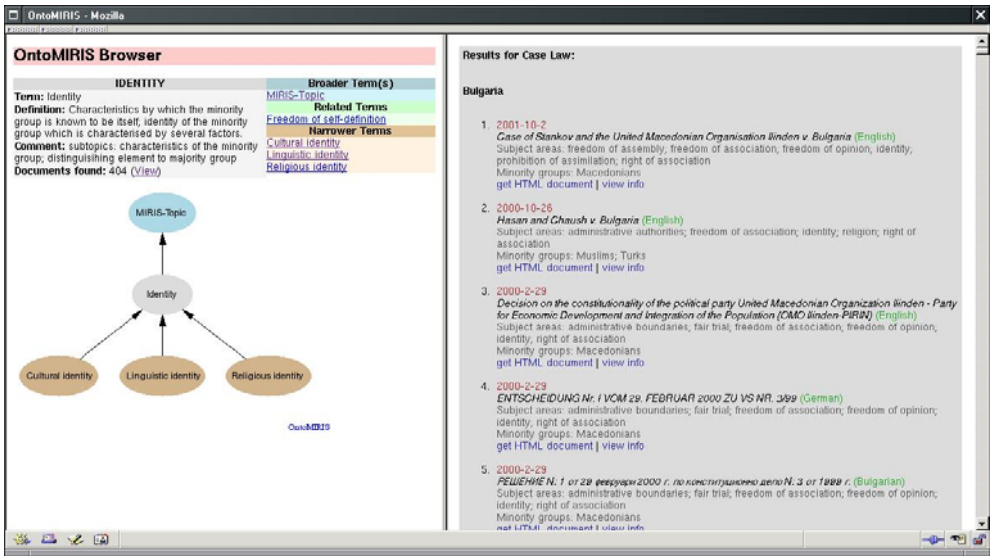



**Grafik 41: Visueller Überblick zur Bearbeitung der Ontologierelationen**




In der in Grafik 40 dargestellten Oberfläche kommt man durch Anklicken der Konzepte auf die Definition, Synonyme, Kommentare und Beziehungen zu anderen Knoten. In der Ansicht von Grafik 41 ist die Hierarchie auf der linken Seite als horizontale Ordner angeordnet, auf der rechten Seite als Graph. Wenn ein Themenknoten ausgewählt wird, dann wird die Anzahl der Treffer und eine Trefferliste gezeigt (Grafik 42).

**Grafik 42: Suchergebnisse im Browserfenster der Ontologieschnittstelle**




Die erste produzierte Ontologie ergab sieben Hierarchieebenen, sie war aber trotzdem zu breit (mehrere Meter als Papierausdruck), um übersichtlich sein zu können.[342] Deshalb wurde die Anzahl der Knoten insgesamt verringert und insbesondere im oberen Teil der Ontologie wurden Konzepte zu abstrakteren Konzepten zusammengefasst, die dann an einer der 10 Hierarchiestufen differenziert werden.[343]

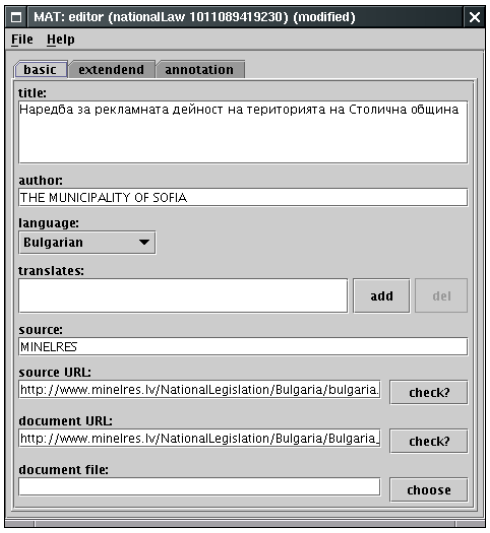
### 7. Zuordnung der Dokumente zu den Ontologieknoten

Wenn eine Entscheidung über die Ordnungsstruktur getroffen wurde, dann müssen die Daten noch dieser Struktur geordnet werden, damit das den Daten implizite Wissen durch die Struktur explizit wird. Bei einfachen oder formalen Ordnungsstrukturen kann dies der Computer übernehmen. Ansonsten verwendet man dazu das Werkzeug, mit dem Dokumente gespeichert und seine formalen Metadaten (Art und Name des Dokumentes usw.) eingegeben und bearbeitet werden. Zusätzlich wird nun die thematische Zuordnung abfragt. Dazu kann in der Ontologie navigiert werden und ein oder mehrere Themen vergeben werden (Grafiken 43 und 44).

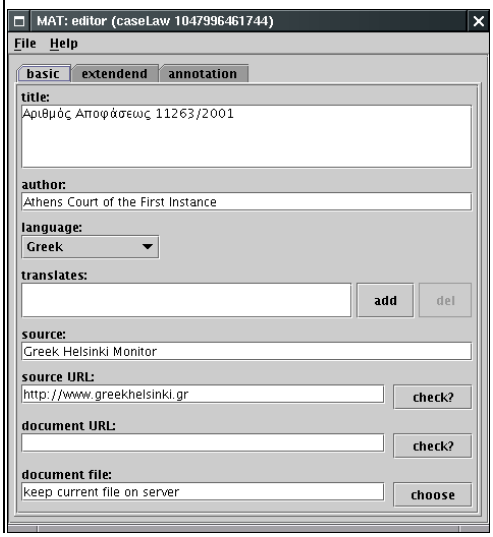
[342] http://dev.eurac.edu/~muser/ontomiris/ontoMIRIS-old.png : 2.3.2004.

[343] Die erstellte MIRIS-Ontologie: http://dev.eurac.edu/~muser/ontomiris/ontoMIRIS-new.png und Hinweise dazu in http://dev.eurac.edu/~muser/ontomiris/: 2.4.2004.

**Grafik 43: Gemeindesatzung von Sofia**




**Grafik 44: Erstinstanzliches Urteil eines Athener Gerichts**




Ein grundsätzliches Problem bei der Erstellung einer Ontologie für eine Dokumentsammlung ist, dass Dokumente nicht identisch sind mit dem Wissen, das sie enthalten. Das Urteil des Athener Gerichts enthält neben den minderheitenrelevanten Erwägungen sicher auch anderes Wissen, z.B. eine Kostenentscheidung. Außerdem kann es sein, dass es sowohl um Sprachenrechte wie um religiöse Rechte geht. Die Dokumente sind also nicht so zugeschnitten, wie das die Wissenseinheiten sein müssen, die in einer Ontologie repräsentiert werden. Die Zuordnung von Dokumenten zu einzelnen Klassen der Ontologie ist daher keine Zuordnung von Wissen, sondern ein sog. *mapping*, eine Zuordnung verschiedener Wissensrepräsentationsklassen zueinander. Auch wenn die MIRIS-Ontologie also perfekt an das rechtliche Wissen angepasst wäre, gäbe es immer Abweichungen zur Wissensrepräsentation durch die Dokumente.

Konkret muss bei der Zuordnung Dokument-Ontologie also in jedem Dokument zunächst das enthaltene Wissen erkannt werden, um es dann in eine der Klassen der Ontologie zu übersetzen. Wegen diesem Zwischenschritt entspricht kaum ein Dokument genau einer Klasse in der Ontologie, sondern repräsentiert stets Wissen zu mehreren Themen. Wünschenswert wäre natürlich eine möglichst scharfe Trennung verschiedener Themen durch eine tiefe Annotierung des Inhalts. Demnach müssten einzelne Sätze in viele Teile zerlegt werden. Der Art. 2 des italienischen Gesetzes zum Schutz der historischen Minderheiten[344] würde zumindest in zwölf Teile für die zwölf genannten Minderheiten zerfallen: "*In attuazione dell'articolo 6 della Costituzione e in armonia con i princípi generali stabiliti dagli organismi europei e internazionali, la Repubblica tutela la lingua e la cultura delle popolazioni albanesi, catalane, germaniche, greche, slovene e croate e di quelle parlanti il francese, il franco-provenzale, il friulano, il ladino, l'occitano e il sardo.*" Der erste Teil des Satzes würde hingegen allen zwölf Klassen zugeordnet werden.

Dieses konkrete Beispiel soll zeigen, dass es leichter ist, Dokumente zu ordnen statt das in ihnen enthaltene Wissen. Umgekehrt ist es leichter, das Wissen aus der Ontologie abzulesen statt aus den Dokumenten. Die Ontologie nimmt also den Teil der Sinnerkennung vorweg und ist deswegen äußerst komplex und arbeitsaufwändig. Nach dem Auslaufen der Finanzierung wurde die Zuordnung von Dokumenten zur Ontologie eingestellt.

[344] Gesetz Nr. 482 v. 15.12.1999.

### 8. Ergebnis und Bewertung der Ontologie für MIRIS

Wahrscheinlich hätte der hohe Arbeitsaufwand die MIRIS-Ontologie aber nicht zum Scheitern gebracht. Es lag letztlich daran, dass das Ergebnis der Zuordnung unter den juristischen Mitarbeitern nicht die Übereinstimmung fand, die nötig gewesen wäre, um Schlüsse aus den expliziten Zusammenhängen ziehen zu können. Drei Normen zum selben Thema, etwa zum Grundschulunterricht in der Muttersprache, müssten nach der Ontologie eine enge Beziehung zueinander haben. Wenn eine Norm jedoch eine verbindliche, aber (noch) nicht umgesetzte Völkerrechtsnorm ist, die zweite eine unverbindliche Völkerrechtsnorm und die dritte eine nicht (mehr) gültige staatliche Norm, dann haben sie eben nur ihr Thema gemeinsam, sie täuschen die Benutzer aber über das scheinbar enthaltene Wissen über eine Rechtssituation. Die Rechtssituation kann seinerseits nur in einer umfassenden und konkreten Fallbetrachtung von ausgesprochenen Spezialisten angegeben werden, denn dazu ist weiteres Wissen etwa über die Norminterpretation notwendig.[345] Der rechtsvergleichende Anwender kann sich also gerade nicht auf die Einteilung in der Ontologie verlassen und daraus das rechtliche Wissen herausziehen, sondern er muss sich den Gesamtblick auf das jeweilige Rechtssystem und seine Zusammenhänge erhalten. Die Einteilung in Themen mag also helfen, sich einen Überblick über Anzahl und Art der möglicherweise einschlägigen Rechtsdokumente zu beschaffen, aber an deduktive Schlussfolgerungen ist nicht zu denken.

Grundsätzlich wäre dieses Problem mit einer Ontologie zu lösen. Es müssten Eigenschaften wie ‚gültig', ‚anwendbar' usw. eingeführt werden. Damit wird das System aber immer unübersichtlicher und weniger verständlich.

Die Dokumente in MIRIS wurden nie der erstellten Ontologie zugeordnet, weil die nachträgliche Annotierung mehrerer tausend Dokumente mehr Zeit und Aufwand bedeutet hätten, als man an Nutzen erwarten konnte. Andererseits, wenn die Ontologie von Anfang an in den Dokumentsammlungsprozess eingebunden gewesen wäre, hätte sie gemeinsam mit der Dokumentsammlung wachsen müssen, was bei jeder Revision einer Klasse die Neuordnung aller ihrer Elemente mit sich gebracht hätte. Das hätte unter dem Strich eher mehr Arbeit bedeutet als weniger.

Ein grundsätzlichere Kritik gegen Ontologien bringt Voß vor: Ontologien seien wie alle Begriffssysteme notwendigerweise lückenhaft und subjektiv, weil sie implizite Angaben über die Welt machen müssen, deren Teil sie selbst sind. Sie müssten also die Fähigkeit haben, die Welt und sich selbst als Teil der Welt zu beschreiben. Die Ontologie enthielte dann sich selbst und die Beschreibung der Welt, was ein Widerspruch ist. Daher kann kein Begriffssystem vollständig sein.[346]

Meiner Ansicht nach liegt ein weiteres Grundproblem darin, dass Wissensordnungssysteme der Dynamik des zugrunde liegenden Wissens gerecht werden müssen. Wandelt sich das Wissen oder ist es sehr stark kontextgebunden, dann müsste das Wissensordnungssystem automatisch nachgeführt werden oder zumindest die Wissensänderung in das Ordnungssystem überführt werden. Keines der vorgestellten Wissensordnungssysteme enthält eine dynamische Komponente, einige kommen jedoch leichter mit Wandel zurecht, weil sie einfach neu aufgebaut werden können. Es handelt

[345] Völkerrechtliche Normen treten regelmäßig nicht mit Unterzeichnung, Ratifizierung oder Hinterlegung der Unterschrift in Kraft, sondern erst dann, wenn eine hinreichende Zahl von ratifizierten Unterschriften hinterlegt und dies den Vertragsstaaten mitgeteilt wurde. Das bedeutet, dass die Rechtsgültigkeit des Vertrages in einem Rechtssystem vom Geschehen in anderen Rechtssystemen abhängt, so dass nicht einmal Spezialisten eines Rechtssystems anhand des normativen Dokuments seine Gültigkeit feststellen können. Ähnliche Beispiele souveränitätsüberschreitender Sachverhalte behandelt das internationale Privatrecht in der Zusammenschau der Regeln mehrerer Rechtssysteme. Im Zusammenhang mit nationalen Minderheiten wird oft die Staatsangehörigkeitsregelung bedeutsam, die auf die internen Regeln anderer Rechtssysteme abstellt, so dass sich der tatsächliche Regelungsgehalt mit der Änderung fremden Rechts selbst ändert: Art. 16, Abs. 1, S. 2 GG: „Der Verlust der Staatsangehörigkeit darf [...] gegen den Willen des Betroffenen nur dann eintreten, wenn der Betroffene dadurch nicht staatenlos wird." Die Bundesrepublik Deutschland macht hier interne Rechtsfolgen davon abhängig, ob zum Zeitpunkt des Entzugs ein anderer Staat der betroffenen Person seine Staatsbürgerschaft zuerkennt. Das führt in z.T. ungelöste Fragen fremder Rechtssysteme.

[346] Voß a.a.O. S. 8.

sich um die automatisch erstellten Schlagwörter, Cluster und Indizes. Gerade weil rechtliches Wissen dynamisch, komplex und kontextabhängig ist, eignen sich Wissensordnungen, die leicht neu aufgebaut werden können und das Interpretieren von Komplexität und Kontext dem Benutzer überlassen.

Als Haupterkenntnis für die Ersteller und Benutzer von Rechtsdatenbanken bleibt, dass der verwendeten Wissensordnung überragende Bedeutung zukommt. Die Tabelle 16 verschafft hier eine Orientierung über die Mächtigkeit und Aufwändigkeit des Ordnungssystems.

# Literaturangaben zu Kapitel 5

Arampatzis A., van der Weide Th. P., Koster C. H. A., Bommel P. van (2000), An Evaluation of Linguistically-motivated Indexing Schemes, S. 34-45 in: Proceedings of the 22nd Annual Colloquium of the British Computer Society Information Retrieval Specialist Group (BCS-IRSG) 2000, Sidney Sussex College Cambridge 2000, http://citeseer.nj.nec.com/arampatzis00evaluation.html : 5.10.2004.

Barzilay R., Elhadad M. (1997), Using lexical chains for text summarization, S. 10-17 in: Proceedings of the Intelligent Scalable Text Summarization Workshop (ISTS) 1997, ACL Madrid 1997, http://citeseer.nj.nec.com/barzilay97using.html : 5.10.2004.

Brunn M., Chali Y., Pichak C. J. (2001), Text Summarization using lexical Chains, S. 135-140 in: Proceedings of the Document Understanding Conference (DUC) 2001, New Orleans 2001, http://citeseer.nj.nec.com/brunn01text.html : 5.10.2004.

Carstensen K.-U. (2004), Nicht-sprachliches Wissen, S. 448-454 in: Carstensen K.-U., Ebert C., Endriss C., Jekat S., Klabunde R., Langer H. (Hrsgg.) (2004), Computerlinguistik und Sprachtechnologie – Eine Einführung, Spektrum Verlag Heidelberg 2004.

Deerwester S., Dumais S. T., Landauer T. K., Furnas G. W., Harshman R. A. (1990), Indexing by latent semantic analysis, S. 391-407 in: Journal of the Society for Information Science 41(6).

DIN 2331, Begriffssysteme und ihre Darstellung, 1980.
DIN 2338, Begriffssystem Zeichen, Teil 1 Der Zeichenbegriff, 1984.
DIN 2342, Begriffe der Terminologielehre, 1992.
DIN 1463 Teil 1: Erstellung und Weiterentwicklung von Thesauri. Einsprachige Thesauri, 1988.
DIN 31623, Indexierung zur inhaltlichen Erschließung von Dokumenten, 1978,
DIN 31623, Teil 3: Syntaktische Indexierung mit Deskriptoren, 1988.
Alle Beuth Verlag Berlin.

Dongilli P., Gamper J. (2003), MIRIS Unleashed, S. 131-139 in: Proceedings JURIX 2003: The Sixteenth Annual Conference on Legal Knowledge and Information Systems December 2003, IOS Press Utrecht 2003.

Englberger H. (1995), Computergestützte Informationsvisualisierung - Eine Klassifikation aktueller Techniken und ihre Einsatzpotentiale für die Unternehmung, Diplomarbeit TU München 1995, http://www11.informatik.tu-muenchen.de/publications/pdf/da-englberger1995.pdf : 5.10.2004.

Europäisches Komitee für Normung, Workshop Agreement CWA 13874, März 2000, http://www.cenorm.be: 5.10.2004.

Glück H. (2000), Metzler Lexikon Sprache, 2. Auflage J. B. Metzler Verlag Stuttgart Weimar 2000.

Gruber T. R. (1993), Towards Principles for the Design of Ontologies Used for Knowledge Sharing, in: Guarino N., Poli R. (Hrsgg.) (1993), Formal Ontology in Conceptual Analysis and Knowledge Representation, Kluwer Academic Publishers Deventer (NL) 1993.

Hansen M. (1998), Möglichkeiten des Einsatzes von Concept-Maps zum juristischen Lernen, in: JurPC Web-Dok. 103/1998.

Hovy E. H. (2003), Text Summarization, S. 599-615 in: The Oxford Handbook of Computational Linguistics, Mitkov R. (Hrsg.) (2003), Oxford University Press Oxford 2003.

ISO 1087 - Terminology / Vocabulary, 1990.

ISO 2788, 1986: 2.

ISO/IEC 13250, 2002 (E).

ISO 15836-2003, Februar 2003: http://www.niso.org/international/SC4/n515.pdf

Kittredge R. I. (2003), Sublanguages and controlled languages, S. 430-447 in: The Oxford Handbook of Computational Linguistics, Mitkov R. (Hrsg.) (2003), Oxford University Press Oxford 2003.

NISO Standard Z39.85-2001 (September 2001): http://www.niso.org/standards/resources/Z39-85.pdf

Nohr H. (2001), Automatische Indexierung, Einführung in betriebliche Verfahren, System und Anwendungen, Materialien zur Information und Dokumentation, Verlag für Berlin-Brandenburg Potsdam 2001.

Noy N. F., Sintek M., Decker S., Crubezy M., Fergerson R. W., Musen M. A. (2001), Creating Semantic Web Contents with Protégé-2000., S. 60-71 in: IEEE Intelligent Systems 16(2):60-71, 2001,

http://www-smi.stanford.edu/pubs/SMI_Reports/SMI-2001-0872.pdf : 5.10.2004.

Runte M., Beißwenger M., Storrer A. (2003), Modellierung eines Terminologienetzes für das automatische Linking auf der Grundlage von WordNet, S. 95-104 in: Kunze C., Lemnitzer L., Wagner A. (Hrsgg.) (2003), Anwendungen des deutschen Wortnetzes in Theorie und Praxis, GermaNet Workshop Proceedings, Tübingen 2003,

http://www.sfs.uni-tuebingen.de/ lsd/GermaNet-Workshop/TermNet-LDV.pdf : 5.10.2004.

Schmidt G. (2002), XBRL ist... das, was die Finanzwelt braucht!, in: Consultant - Steuern Wirtschaft Finanzen, 05/2002, Max Schimmel Verlag Würzburg 2002.

Shannon, C. E. (1998), The Mathematical Theory of Communication, University of Illinois Press Urbana/ Chicago 1998.

Shapiro S.C. (1992), Artificial Intelligence, S. 54-57 in: S. C. Shapiro, (Hrsg.) (1992), Encyclopedia of Artificial Intelligence, 2. Aufl. John Wiley & Sons Inc. New York 1992.

Silber K.F., McCoy H.G. (2000), Efficient Text Summarization Using Lexical Chains, S. 252-255 in: Proceedings of the 5th international conference on Intelligent user interfaces, New Orleans 2000, http://web.media.mit.edu/~lieber/IUI/Silber/Silber.pdf: 5.10.2004.

Teich E., Fankhauser P. (2004), WordNet for Lexical Cohesion Analysis, S. 326-331 in: Sojka P., Pala K., Smrž P., Fellbaum C., Vossen P. (Hrsgg.) (2004), Proceedings of the Second International WordNet Conference (GWC 2004), Masaryk Universität Brno 2003.

Umstätter W. (1992), Nutzen der Indexierung bei Online-Datenbanken, S. 403-420 in: 14. Online-Tagung 1992 Frankfurt am Main, DGD-Schrift (OLBG-13) 2/92, http://www.ib.hu-berlin.de/~wumsta/pub65.html: 17.11.2003.

Vossen P. (2003), Ontologies, S. 464-482 in: The Oxford Handbook of Computational Linguistics, Mitkov R. (Hrsg.) (2003), Oxford University Press Oxford 2003.

Voß J. (2004), Begriffssysteme – Ein Vergleich verschiedener Arten von Begriffssystemen und Entwurf des integrierenden Thema-Datenmodells, Studienarbeit im Diplomstudiengang Informatik, HU Berlin, Version 1.1d, http://www.nichtich.de/epub/begriffssysteme03/begriffssysteme.pdf : 8.3.2004.

Internetquellen in Zitierreihenfolge:
http://ai.kaist.ac.kr/~sjcho/semantic-web/ : 2.3.2004.
http://www.toolexpert.de/gif/gif_catia_kbe/2.jpg : 28.10.2003.
http://www.phil-fak.uni-duesseldorf.de/~rumpf/talks/automatentheorie.pdf: 10.11.2003.
http://www.cyc.com/cyc/technology/whatiscyc_dir/whatsincyc : 15.3.2004.
http://www.ericfacility.net/extra/pub/ialsearch.cfm : 28.10.2003
http://www.tessmann.it/seiten/opac/index_einfach_de.html : 28.10.2003
http://www.jura.uni-duesseldorf.de/rave/e/ravesyse.htm : 28.10.2003
http://vivisimo.com: 17.11.2003.
http://www.cs.uni-bonn.de/~arbuckle/abis/semi-automatic_results.html : 28.10.2003
http://www.dbs.informatik.uni-muenchen.de/~kailing/Fopras/singleLink.html : 28.10.2003
http://www.techquila.com/topicmaps/tmworld/ : 28.10.2003
http://www.lexisnexis.com/infopro/products/index/thesABnew.shtml: 17.11.2003,
http://bigmac.phil.uni-sb.de/trex: 17.11.2003.
http://www.uni-saarland.de/fak5/ezw/abteil/lehr/concept/concept-map : 28.10.2003
http://www.december.com/web/text/images/cyberland.gif : 28.10.2003.

http://www.elama.de/assoc : 23.9.2004.
http://www.museumsbund.de/fgdoku/dmbdoku_termine/dmbokt2002/beitraege/normdaten_junger.pdf: 14.11.2003.
http://www.wsiat.on.ca: 17.11.2003.
http://www.wsiat.on.ca/ExtDec/KeywordDirectory.asp: 17.11.2003.
http://archiv.tu-chemnitz.de/pub/2002/0127: 4.11.2003.
http://www.itl.nist.gov/iaui/894.02/related_projects/tipster_summac: 17.11.2003.
http://citeseer.nj.nec.com/591246.html: 17.11.2003.
http://extranet.staedtetag.de/suche/index.html: 17.11.2003.
http://bibliothek.bbaw.de/Goethe/my_html/wortschatz.htm: 17.11.2003.
http://is.uni-sb.de/studium/handbuch/exkurs.ind.php#intellind: 19.11.2003.
http://www.yahoo.com: 18.11.2003.
http://www.public.iastate.edu/~CYBERSTACKS/CTW.htm: 17.11.2003.
http://www.lub.lu.se/desire/sbigs.html: 17.11.2003.
http://bubl.ac.uk: 17.11.2003.
http://muchmore.dfki.de/ : 23.9.2004.
http://javelina.cet.middlebury.edu/lsa/out/tutorial.htm : 28.10.2003.
http://www.biologie.uni-hamburg.de/b-online/e43/t1.htm : 2.2.2004.
http://www.xbrl-deutschland.de: 19.11.2003.
http://www.xbrl-deutschland.de/GermanAP_2002_02_15_nav.htm: 19.11.2003.
http://www.cs.uni-bonn.de/~arbuckle/abis/semi-automatic_results.html: 19.11.2003.
http://www.iuk.hdm-stuttgart.de/nohr/Km/KmPubl/wisska/wisska_1.html: 19.11.2003.
http://www.vtt.fi/ele/research/soh/projects/totem2001/kmpmhtml/kmpm4.1.2.html : 3.2.2004.
http://www.ontopia.net/download/freedownload.html : 12.3.2004.
http://www.y12.doe.gov/sgml/sc34/document/0129.pdf : 2.2.2004.
http://www.willpower.demon.co.uk/thessoft.htm : 11.3.2004.
http://www.iim.fh-koeln.de/dtp/werkzeuge.html : 11.9.2004.
http://www.dicomaker.netfirms.com/ : 11.3.2004. http://www.lai.com/lai/tg.html : 11.3.2004.
http://www.learn-line.nrw.de/angebote/selma/foyer/projekte/hammproj4/difference.htm : 3.2.2004.
http://www.mindmap.ch/software.htm : 11.3.2004.
http://www.denkzeichnen.de/: 3.2.2004.
http://beat.doebe.li/bibliothek/w00085.html : 3.2.2004.
http://infolab.uvt.nl/prj/macs/macsd.html : 15.3.2004.
http://www.jura.uni-sb.de/internet/gericht.html : 5.2.2004.
http://www.bundesverfassungsgericht.de/cgi-bin/link.pl?entscheidungen : 5.2.2004.
http://www.bundesgerichtshof.de/entscheidungen/entscheidungen.php : 5.2.2004.
http://www.bverwg.de/enid/926d526ba0f8166e40d164301a1aa626,0/96.html : 5.2.2004.
http://bundesrecht.juris.de : 5.2.2004.
http://www.lgkassel.de/ : 5.2.2004.
http://www.eurac.edu/miris : 24.2.2004.
http://www.minelres.lv/ : 24.2.2004.

http://www.greekhelsinki.gr/bhr/english/index.html : 24.2.2004.
http://www.ciemen.org/mercator/index-gb.htm : 24.2.2004.
http://dublincore.org/documents/dces/ : 24.2.2004.
http://www.cogsci.princeton.edu/~wn/ : 25.2.2004.
http://www.cogsci.princeton.edu/cgi-bin/webwn : 25.2.2004.
http://protege.stanford.edu/ : 24.2.2004.
http://dev.eurac.edu/~muser/ontomiris/ontoMIRIS-new.png : 2.3.2004.
http://dev.eurac.edu/~muser/ontomiris/: 2.4.2004.

# Fazit

Welche computerlinguistischen Methoden bieten sich zur Untersuchung des auf Schriftlichkeit ausgelegten Rechts an? Welche Erkenntnisse konnten in dieser Dissertation gewonnen werden?

Im Einklang mit der Annahme, dass computerlinguistische Methoden für Rechtsdatenbanken einen einfach zu realisierenden Vorteil bieten müssen, wird in Kapitel 1 wird eine Methode vorgestellt, wie ein Rechtskorpus aufgebaut werden kann. Der Schritt zur korpuslinguistischen Ausnutzung des Rechtskorpus ist damit vorgezeichnet. Die Methode zur Erlangung und Klassifizierung rechtlicher Texte, die für sich genommen von untergeordneter Bedeutung zu sein scheint, kann durch ihre **Katalysatorwirkung** viel zur **Heranführung der Rechtswissenschaft an computerlinguistische Verfahren** leisten. Auch die traditionellsten Juristen erkennen mittlerweile an, dass das Internet für ihre Arbeit relevante Informationen bereitstellt, und wenn es nur das Formular für einen elektronischen Mahnbescheid ist. Damit liegt die Motivation zum Speichern dieser Informationen, zum Erstellen eines Rechtskorpus, auf der Hand.

Im Kapitel 2 wird die Problematik des Konzepts ‚Fachsprache' aufgezeigt. Die Bezeichnung eines Textes als ‚fachsprachlich' ist letztlich das Ergebnis einer wertenden Gesamtschau von Lexikon, Grammatik und Kontext (Sender, Empfänger, Medium, Textausschnitt). Trotzdem gelingt es, mit zwei relativ einfachen Verfahren, die Fachsprachlichkeit eines Textes automatisch zu erkennen, bzw. die zu einem gewissen Grad subjektiven Urteile von Terminologen derart gut zu reproduzieren, dass **bei hinreichender Textlänge** die **Fachgebietserkennung** als **gelöstes Problem** bezeichnet werden kann. In einem Vorversuch konnte sogar die nur implizit eingeflossene Hierarchie der Fachgebiete errechnet werden. Die praktische Bedeutung der vorgestellten oder ähnlichen Methoden für Terminologie- und Terminografie auch anderer Fachsprachen dürfte erheblich sein.

Die in Kapitel 3 theoretisch und technisch beschriebene Methode dynamischer Termdarstellung hat sich bereits in großen Terminografieprojekten bewährt. Das Unterteilen einer komplexen Aufgabe in einzelne Modulbausteine ist eine in der Informatik allgemein anerkannte Strategie und scheint den bisherigen Ansätzen einer untrennbaren Einheit überlegen zu sein. Die konkret vorgestellte Kombination der Module (‚Grammatiken') zur Erzeugung eines Mensch-Maschine-Dialogs hingegen soll nur ein anschauliches Beispiel implementieren. Der Dialogpartner Computer hat hier noch viel zu lernen (sehen, hören, mitdenken, antizipieren, Hintergrundwissen sammeln) und es wird ihm in rasender Geschwindigkeit beigebracht. Die **Terminografie** sollte sich dieser Entwicklung öffnen, indem sie **analytisch-synthetisch** vorgeht und die Fokussierung auf ein bestimmtes Endprodukt aufgibt. Kapitel 3 **beschreibt die Vorteile und weist ihre prinzipielle Realisierbarkeit nach**.

In Kapitel 4 wird der voraussetzungsarme computerlinguistische Ansatz mit Beispieltermen auf zwei prominente Aufgaben in Rechtsdatenbanken angewandt, die Termextraktion für Termdatenbanken und die Indizierung für Textdatenbanken. Der Ansatz erweist sich als echte Alternative zur Konkurrenz, dem linguistischen und dem statistischen Ansatz. **Der beispielbasierte Ansatz** wurde hier besprochen, weil er sich besonders **für Umsteiger auf computerlinguistische Methoden** eignet, viel Effekt für geringen Aufwand verspricht und bei klassischen, praktisch in jeder Rechtsdatenbank anfallenden Problemstellungen eingesetzt werden kann. Darüber hinaus lässt eine Kombination des beispielbasierten Ansatzes mit anderen Ansätzen eine Ergebnisverbesserung erwarten, so dass die vorgestellte Methode auch für Projekte von Interesse sein dürfte, die bereits computerlinguistische Methoden einsetzen.

Die unmittelbar mit der Klassifizierung zusammenhängende Frage nach der Ordnung rechtlichen Wissens wird in Kapitel 5 prozessual beantwortet. Nicht eine Methode der Ordnung, sondern die Vorgehensweise beim Auffinden der am besten geeigneten Methode wird aufgezeigt. Allein aus der Zusammenstellung der **Vielzahl an Methoden** mit ihren jeweiligen Stärken und Anwendungsgebieten macht deutlich, dass hier eine Entscheidung weit überlegter erfolgen muss, als das erfahrungsgemäß oft geschieht. So sehr in den anderen Kapiteln aufgefordert worden ist, den computerlinguistischen Methoden zu vertrauen, so sehr soll hier vor einer vorschnellen Entscheidung für eine bestimmte Methode gewarnt werden. Das Kapitel 5 versucht daher, die komplexe Aufgabe von Wissensordnungen theoretisch so weit aufzuarbeiten, dass die Verantwortlichen von Rechtsdatenbanken daraus **das für eine verantwortliche Entscheidung notwendige technische Wissen** ziehen können.

Die Arbeit gewährt damit einen fundierten Einblick in aktuelle computerlinguistische Methoden für Rechtsdatenbanken.

# Gesamtbibliografie

Aha D.W. (1997), Editorial-lazy learning, S. 1-3 in: Artificial Intelligence Review 1997/11.

Ahmad K., Davies A. E. (1994), Weirdness in Special-language Text: Welsh Radioactive Chemicals Texts as an Exemplar, S. 22-52 in: Veröffentlichungsreihe des Internationalen Instituts für Terminologieforschung (IITF-Series) 1994, Band 5, Nr. 2, TermNet Wien 1994.

Apresjan J. D., Boguslavskij I. M., Iomdin L. L., Lazurskij A. V., Sannikov V. Z., Tsinman L. L. (1992), ETAP-2: The Linguistics of a Machine Translation System, S. 97-112 in: Meta, Bd. 37:1, http://www.erudit.org/revue/meta/1992/v37/n1/001895ar.pdf

Arampatzis A., van der Weide Th. P., Koster C. H. A., Bommel P. van (2000), An Evaluation of Linguistically-motivated Indexing Schemes, S. 34-45 in: Proceedings of the 22nd Annual Colloquium of the British Computer Society Information Retrieval Specialist Group (BCS-IRSG) 2000, Sidney Sussex College Cambridge 2000, http://citeseer.nj.nec.com/arampatzis00evaluation.html : 5.10.2004.

Ardissono L., Goy A., Petrone G., Segnan M., Torasso P. (2003), Intrigue: Personalized Recommendation Of Tourist Attractions For Desktop And Handset Devices, S. 687-714 in: Applied Artificial Intelligence: Special Issue on Artificial Intelligence for Cultural Heritage and Digital Libraries (2003) 17:8-9, http://www.di.unito.it/~liliana/EC/aai03.pdf : 29.3.2004.

Arntz R., Picht H., Mayer F. (2002), Einführung in die Terminologiearbeit, Studien zu Sprache und Technik Bd. 2, Georg Olms Verlag Hildesheim 2002.

Arppe A. (1995), Term extraction from unrestricted text, short paper in: Koskenniemi K. (Hrsg.) (1995), Proceedings of the 10th Nordic Conference of Computational Linguistics (Nordiska datalingvistdagarna: NoDaLiDa 1995), Helsinki 1995, http://www.lingsoft.fi/doc/nptool/term-extraction.html : 5.10.2004.

Banko M., Brill E. (2001), Scaling to Very Very Large Corpora for Naturale Language Disambiguation, S. 26-33 in: Meeting of the Association for Computational Linguistics 2001, http://citeseer.nj.nec.com/banko01scaling.html : 18.6.2004.

Barzilay R., Elhadad M. (1997), Using lexical chains for text summarization, S. 10-17 in: Proceedings of the Intelligent Scalable Text Summarization Workshop (ISTS) 1997, ACL Madrid 1997, http://citeseer.nj.nec.com/barzilay97using.html : 5.10.2004.

Beißwenger M., Storrer A., Runte M. (2003), Modellierung eines Terminologienetzes für das automatische Linking auf der Grundlage von WordNet, S. 95-104 in: Kunze C., Lemnitzer L., Wagner A. (Hrsgg.)(2003), Tagungsband des LDV-Forum 19 (2003): 1-2, Sonderheft mit Beiträgen des GermaNet Workshops zu Anwendungen des deutschen Wortnetzes in Theorie und Praxis.

Bergenholtz H., Tarp S. (Hrsg.) (1995), Manual of Specialised Lexicography – The preparation of specialized dictionaries, John Benjamins Amsterdam 1995.

Bonnet E., Gaussier E., Langé J.-M. (1994), A method for automatic extraction of terms from bilingual corpora, in: Proceedings of the 14th International Conference on Artificial Intelligence KBS, Expert Systems and Natural Language (AVIGNON-94), EC2 Nanterre 1994.

Bortz J. (1993), Statistik für Sozialwissenschaftler, 4. Auflage Springer Berlin u.a. 1993.

Bourigault D., Jacquemin C. (1999), Term extraction and term clustering. An integrated platform for computer-aided-terminology, S. 15-22 in: Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics (EACL 1999), EACL Bergen 1999, http://citeseer.nj.nec.com/bourigault99term.html : 5.10.2004.

Bourigault D., Jacquemin C., L'Homme M.-C., (Hrsgg.) (2001), Recent Advances in Computational Terminology, John Benjamins Publishing Company Amsterdam 2001.

Brekke M., Myking J., Ahmad K. (1996), Terminology management and lesser-used living languages: A critique of the corpus-based approach, S. 179-189 in: Sandrini P. (Hrsg.) (1996), Proceedings of the 4th International Congress on Terminology and Knowledge Engineering (TKE'96) Innsbruck, TermNet Wien 1996, ftp://ftp.ee.surrey.ac.uk/pub/research/AI/TKE.papers/Postscript_versions/Terminology_Management.ps.gz : 5.10.2004.

Brill E. (1995), Transformation-based error-driven learning and natural language processing: A case study in part of speech tagging, Computational Linguistics, 1995.

Brin, S., Page, L. (1998), The anatomy of a large-scale hypertextual Web search engine, S. 107-117 in: Computer Networks and ISDN Systems Vol. 30, Nr.1-7 1998, http://citeseer.nj.nec.com/brin98anatomy.html: 10.10.2003.

Brinkmann K.-H., Schulz J., Tanke E., (1969), Das Wörterbuch aus der Maschine, S. 9-15 in: Data Report 4/4 1969; nachgedruckt in: Graham J. D., Grewe K., Reisen U. (Hrsgg.) (1995), Terminologiearbeit. Theorie und Praxis, in: Festschrift für Eberhard Tanke zum 75. Geburtstag, Deutscher Terminologie-Tag Köln 1995.

Brockmeier, J. (1999): Die Welt als Bibliothek, Colloquium vom 15.7.1999, http://www.fu-berlin.de/postmoderne-psych/berichte3/brockmeier.htm : 13.11.2003.

Brunn M., Chali Y., Pichak C. J. (2001), Text Summarization using lexical Chains, S. 135-140 in: Proceedings of the Document Understanding Conference (DUC) 2001, New Orleans 2001, http://citeseer.nj.nec.com/brunn01text.html : 5.10.2004.

Budin G. (2002), Der Zugang zu mehrsprachigen terminologischen Ressourcen – Probleme und Lösungsmöglichkeiten, in: In Mayer F., Schmitz K.-D., Zeumer J. (Hrsgg.), eTerminologie - Professionelle Terminologiearbeit im Zeitalter des Internet, Tagungsakte des Symposions "eTerminology", Deutscher Terminologie-Tag (DTT) Köln 2002.

Bullo F., Ciola B., Coluccia S., Maganzi Gioeni D'Angiò F., Mayer F., Treiber A., Voltmer L. (2003), Terminologisches Wörterbuch zum Vertragsrecht: italienisch/deutsch Dizionario terminologico del diritto dei contratti italiano – tedesco, C.H.Beck München, Athesia Bozen, Stämpfli Bern, Linde Wien 2003.

Bungarten, T. und Engberg, J. (Hrsg.), (2003): Recht und Sprache - eine internationale Bibliographie in juristischer und linguistischer Fachsystematik, in: Hamburger Arbeiten zur Fachsprachenforschung 05, Attikon-Verlag Tostedt 2003.

Busse, D. (2002): Juristische Auslegungstätigkeit in linguistischer Hinsicht, S. 136-162 in: Haß-Zumkehr U. (Hrsg.) (2002), Sprache und Recht, Jahrbuch 2001 des Instituts für deutsche Sprache, de Gruyter Verlag Berlin.

Cabré Castellvì M.T., Estopà Bagot R., Palatresi J. V. (2001), Automatic term detection: A review of current systems, S. 53-89 in: Bourigault D., Jacquemin C., L'Homme M. C. (Hrsgg.) (2001), Recent Advances in Computional Terminology, Natural Language Processing Band 2, John Benjamins Publishing Company Amsterdam/ Philadelphia 2001.

Carl M., Schaible J., Pease C. (1998), Enhancing translation memory (TM) technologies with linguistic intelligence, MULTI-DOC Deliverable D 4.1 WP 6, Kommission der Europäischen Gemeinschaften Luxemburg 1998.

Carstensen K.-U. (2004), Nicht-sprachliches Wissen, S. 448-454 in: Carstensen K. U., Ebert C., Endriss C., Jekat S., Klabunde R., Langer H. (Hrsgg.) (2004), Computerlinguistik und Sprachtechnologie – Eine Einführung, Spektrum Verlag Heidelberg 2004.

Carstensen K. U., Ebert C., Endriss C., Jekat S., Klabunde R., Langer H. (Hrsgg.) (2004), Computerlinguistik und Sprachtechnologie – Eine Einführung, Spektrum Verlag Heidelberg 2004.

Chakrabarti S., Dom B. und Indyk P. (1998), Enhanced hypertext categorization using hyperlinks, S. 307-318 in: Haas L. M., Tiwary A. (Hrsgg.) (1998), Proceedings ACM SIGMOD International Conference on Management of Data, ACM Press Seattle 1998.

Charniak, E. (1996), Tree-bank grammars, S. 1031-1036 in: Proceedings of the 13th National Conference on Artificial Intelligence (AAAI-96), American Association for Artificial Intelligence (AAAI) Portland 1997.

Chien L.-F., Chen C.-L. (2001), Incremental extraction of domain-specific terms from online text resources, S. 89-109 in: Bourigault D., Jacquemin C. und L'Homme M.-C., (Hrsgg.) (2001), Recent Advances in Computational Terminology (Natural Language Processing), John Benjamins Amsterdam 2001.

Church W., Hanks P. (1989), Word association norms, mutual information and lexicography, S. 76–83 in: Proceedings of the 27th Annual Meeting of the Annual Meeting of the Association for Computational Linguistics (ACL), ACL Vancouver 1989, http://citeseer.nj.nec.com/church89word.html : 5.10.2004.

Ciola B. (2001), Darstellung von Äquivalenzbeziehungen in der übersetzungsorientierten Terminologiearbeit im Recht, S. 742-752 in: Mayer F. (Hrsg.) (2001), Language for Special Purposes: Perspectives for the New Millennium, Bd. 2, Narr Tübingen 2001.

Cowie, J., Ludovik, E., Zacharski, R. (1998), An autonomous, web-based multilingual corpus collection tool, S. 142-148 in: Proceedings of the International Conference on Natural Language Processing and Industrial Applications (NLP+IA), University of Moncton Moncton 1998.

Daille B. (1995), Combined approach for terminology extraction: lexical statistics and linguistic filtering, S. 515-521 in: Proceedings of the 15th International Conference on Computational Linguistics Kyoto (COLING 94) 1994, ICCL Kyoto 1994, http://acl.ldc.upenn.edu/C/C94/C94-1084.pdf : 4.10.2004.

Daille B., Enguehard C., Jacquin C., Raharinirina R. L., Ralalaoherivony B. S., Lehmann C. (2000), Traitement automatique de la terminologie en langue malgache, S. 225–242 in: Mariani J., Masson N., Néel F., Chibout K. (Hrsgg.) (2000), Ressources et évaluation en ingénierie des langues, Actualités scientifiques - Universités Francophones, De Boek and Larcier Paris 2000.

Damashek M. (1995), Gaugin similarity via n-grams: Language-independent sorting, categorization, and retrieval of text, S. 843-848 in: Science, 1995: 267.

Day D., Aberdeen J., Hirschman L., Kozierok R., Robinson P., Vilain M. (1997), Mixed-initiative development of language processing systems, s. 348-355 in: 5th Conference on Applied Natural Language Processing (ANLP-97), Association for Computational Linguistics, Washington D.C. 1997.

De Carolis B. (1999), Generating Mixed-Initiative Hypertexts: A Reactive Approach, S. 71-78 in: Proceedings of the 1999 International Conference on Intelligent User Interfaces (IUI ’99), IUI San Diego 1999, http://citeseer.ist.psu.edu/decarolis99generating.html :29.3.2004.

De Carolis B., De Rosis F., Grasso F., Rossiello A., Berry D. C., Gillie T. (1996), Generating recipient-centered explanations about drug prescription, S. 123-145 in: Artificial Intelligence in Medicine, Vol. 8 (1996): 2.

De Carolis B., De Rosis F., Pizzutilo S. (1997), Generating user-adapted hypermedia from discourse plans, in: Lenzerini M., (Hrsg.) (1997), Proceedings of the 5th Congress of the Italian Association for Artificial Intelligence (AI*IA97), Lecture Notes in Artificial Intelligence (LNAI) Teilreihe von Lecture Notes in Computer Science (LNCS), 1321, Springer Heidelberg 1997.

De Carolis B., Pizzutilo S., Palmisano I. (2003), D-ME: Personal Interaction in Smart Environments, S. 388-392 in: Lecture Notes in Computer Science, Vol. 2702, Springer-Verlag Heidelberg 2003.

Deerwester S., Dumais S. T., Landauer T. K., Furnas G. W., Harshman R. A. (1990), Indexing by latent semantic analysis, S. 391-407 in: Journal of the Society for Information Science 41(6).

DIN 2331, Begriffssysteme und ihre Darstellung, 1980.
DIN 2338, Begriffssystem Zeichen, Teil 1 Der Zeichenbegriff, 1984.
DIN 2342, Begriffe der Terminologielehre, Deutsches Institut für Normung Berlin 1992.
DIN 1463 Teil 1: Erstellung und Weiterentwicklung von Thesauri. Einsprachige Thesauri, 1988.
DIN 31623, Indexierung zur inhaltlichen Erschließung von Dokumenten, 1978,
DIN 31623, Teil 3: Syntaktische Indexierung mit Deskriptoren, 1988.
Alle Beuth Verlag Berlin.

Dongilli P., Gamper J. (2003), MIRIS Unleashed, S. 131-139 in: Proceedings JURIX 2003: The Sixteenth Annual Conference on Legal Knowledge and Information Systems December 2003, IOS Press Utrecht 2003.

Ekmekçioglu F. Ç., Lynch M. F., Robertson A. M., Sembok T. M. T., Willett P. (1996), Comparison of n-gram matching and stemming for term conflation in English, Malay, and Turkish texts, S. 1-14 in: Text Technology, 6/1996, sowie in: Information Research, Vol. 2 No. 2, Oktober 1996,
http://informationr.net/ir/2-2/paper13.html : 5.10.2004.

Engberg, J. (1993), Prinzipien einer Typologisierung juristischer Texte, S. 31-38 in: Fachsprache 15 1/2, Wien 1993.

Englberger H. (1995), Computergestützte Informationsvisualisierung - Eine Klassifikation aktueller Techniken und ihre Einsatzpotentiale für die Unternehmung, Diplomarbeit TU München 1995, http://www11.informatik.tu-muenchen.de/publications/pdf/da-englberger1995.pdf : 5.10.2004.

Enguehard C., Pantera L. (1994), Automatic natural acquisition of a terminology, S. 27-32 in: Köhler R. (Hrsg.), Journal of Quantitative Linguistics 2(1), International Quantitative Linguistics Association (IQLA), Trier 1994.

Europäisches Komitee für Normung, Workshop Agreement CWA 13874, März 2000, http://www.cenorm.be: 5.10.2004.

Faber P., Lopés Rodriguez C. I. und Tercedor Sánchez M. I. (2002), Utilización de técnicas de corpus en la representatición del conocimiento médico, S. 167-197 in: Terminology, 2002, 7(2).

Find J., Kobsa A. und Nill A., User-oriented adaptivity and adaptability in the AVANTI project, in: Conference "Design for the Web: Empirical Studies, Microsoft, Redmond, WA, 1996.

Firth J. R., Palmer F. (Hrsgg.) (1968), Selected Papers of J.R. Firth 1952-59, Longman's Linguistic Library London 1968.

Frilling S. (1994), Textsorten in juristischen Fachzeitschriften, Internationale Hochschulschriften 138, Waxmannverlag Münster/New York 1995, zugl. Diss. Univ. Münster 1994.

Furuse, O. and Iida, H. (1992), An example-based method for transfer-driven Machine Translation, in: The Third International Conference on Theoretical and Methodological Issues (TMI), Empiristic vs. Rationalist Methods in MT, Canadian Workplace Automation Research Center Montréal 1992.

Gamper J. (1999), Construction of a Parallel Text Corpus Encoding Primary Data, in: Academia Nr: 18 (März - Juni 1999), EURAC, Bozen 1999, http://www.eurac.edu/Press/Academia/18/Art_13.asp :30.3.2004.

Giese H., Kleppin K., Schlagwort ‚Falsche Freunde', S. 204 in: Glück H (2000), Metzler Lexikon Sprache, 2. Auflage J. B. Metzler Verlag Stuttgart Weimar 2000.

Glenn H. P. (2000): Legal Traditions of the World, Oxford University Press New York 2000.

Glück H. (2000), Metzler Lexikon Sprache, 2. Auflage J. B. Metzler Verlag Stuttgart Weimar 2000.

Graham J. D., Grewe K., Reisen U. (Hrsgg.) (1995), Terminologiearbeit. Theorie und Praxis, in: Festschrift für Eberhard Tanke zum 75. Geburtstag, Deutscher Terminologie-Tag Köln 1995.

Grefenstette G. (1995), Comparing Two Language Identification Schemes, S. 263-268 in: Bolasco S., Lebart L., Salem A. (Hrsgg.) (1995), Proceedings of the 3. International Conference on Statistical Analysis of Textual Data (Journées d'Analyse de Données Textuelles JADT 95), CISU Rom 1995,
http://www.xrce.xerox.com/Publications/Attachments/1995-012/Gref---Comparing-two-language-identification-schemes.pdf : 8.4.2004.

Gruber T. R. (1993), Towards Principles for the Design of Ontologies Used for Knowledge Sharing, in Guarino N., Poli R., (Hrsgg.) (1993), Formal Ontology in Conceptual Analysis and Knowledge Representation, Kluwer Academic Publishers Deventer (NL) 1993.

Hahn W. v. (1983), Fachkommunikation: Entwicklung, linguistische Konzepte, betriebliche Beispiele, Sammlung Göschen 2223, de Gruyter Berlin, New York 1983.

Halliday M. A. K. (1994), An Introduction to Functional Grammar, 2. Aufl. Edward Arnold London 1994.

Han E.-H. S. und Karypis G. (2000), Centroid-based document classification: Analysis & experimental results, S. 424-431 in: 2000, URL, http://www.cs.umn.edu/karypis: 10.10.2003.

Hansen M. (1998), Möglichkeiten des Einsatzes von Concept-Maps zum juristischen Lernen, in: JurPC Web-Dok. 103/1998.

Heid U. (1999), Extracting terminologically relevant collocations from German technical Texts, S. 241-255 in: Sandrini P. (Hrsg.) (1999), Proceedings of the 5th International Congress on Terminology and Knowledge Engineering (TKE) 1999, TermNet Wien 1999, http://citeseer.nj.nec.com/heid99extracting.html : 5.10.2004.

Heinrich-Litan L. N. (2003), Exakte L∞-Nächster-Nachbar-Suche in hohen Dimensionen - Exact L∞-Nearest-Neighbor Search in High Dimensions, digitale Dissertation an der FU Berlin, http://www.diss.fu-berlin.de/2003/80/index.html : 25.9.2004.

Henn, H., Sitta, H., Wiegand, H.E. (Hrsg.), (1998), Reihe Germanistische Linguistik 191, Max Niemeyer Verlag Tübingen 1998.

Herberger M., Systematik des Bundesrechts, Projektbericht, Juristisches Internetprojekt Saarbrücken, 1997, http://www.jura.uni-sb.de/BGBI/BGBLSYST.HTML: 10.10.2003.

Heutger, V. (2000): Law and Language in the European Union, in: Global Jurist Topics, Vol. 3, Issue 1, Article 3, S. 4.

Hoffmann L. (1975), Fachsprachen und Sprachstatistik. Beiträge zur angewandten Sprachwissenschaft, Akademieverlag Berlin 1975.

Holmes-Higgin P., and Khurshid A. (1996), Is your Terminology in Safe Hands? Data Analysis, Data Modelling and Term Banks, Terminology and Knowledge Engineering, S. 215-224 in: Proceedings of the 4th International Congress on Terminology and Knowledge Engineering Vienna (TKE ’96), INDEKS-Verlag Frankfurt 1996.

Holzborn T. (2003), Die Geschichte der Gesetzespublikation, Tenea Verlag Berlin 2003.

Hong M., Fissaha S., Haller J. (2001), Hybrid filtering for extraction of term candidates from German technical texts, Poster in: Proceedings of Terminologie et Intelligence Artificielle (TIA), Institut National de l’Information Scientifique et Technique (INIST) Nancy 2001,

http://www.iai.uni-sb.de/docs/term_extract.pdf : 5.10.2004.

Honsell, H. (2001): Naturrecht und Positivismus im Spiegel der Geschichte, S. 593-602 in: Festschrift Koppensteiner (2001), Orac Verlag Wien 2001.

Hovy E. H. (2003), Text Summarization, S. 599-615 in: The Oxford Handbook of Computational Linguistics, Mitkov R. (Hrsg.) (2003), Oxford University Press Oxford 2003.

Hovy E. H. (1993), Automated Discourse Generation Using Discourse Structure Relations, S. 341-385 in: Artificial Intelligence (AI) 63 (1993): 1-2,
http://citeseer.ist.psu.edu/hovy93automated.html :29.3.2004.

Humphreys K. (1997), Formalising Pragmatic Information for Natural Language Generation, Dissertation University of Edinburgh, Centre for Cognitive Science, 1995.
http://citeseer.ist.psu.edu/humphreys97formalising.html :30.3.2004.

ISO 1087 - Terminology / Vocabulary, 1990.
ISO 2788, 1986: 2.
ISO/IEC 13250, 2002 (E).
ISO 15836-2003, Februar 2003: http://www.niso.org/international/SC4/n515.pdf

Jacquemin C. (1999), Syntagmatic and paradigmatic representation of term variation, S. 341-348 in: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL) 1999, ACL Maryland 1999,
http://citeseer.nj.nec.com/jacquemin99syntagmatic.html : 5.10.2004.

Jacquemin C., Bourigault D. (2003), Term Extraction and Automatic Indexing, in: Handbook of Computational Linguistics, Mitkov R. (Hrsg.) (2003), Oxford University Press Oxford 2003.

Jain A. K., Duin R. P. und Mao J. (2000), Statistical pattern recognition: A review, S. 4-37 in: IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(1).

Kang H.J., Doermann D. (2003), Evaluation of the Information Theoretic Construction of Multiple Classifier Systems, Proceedings of the 7th International Conference on Document Analysis and Recognition (ICDAR 2003), IEEE Edinburgh 2003, http://citeseer.ist.psu.edu/ps/698712 : 26.9.2004.

Kienast B. (1994), Die Altorientalischen Codices zwischen Mündlichkeit und Schriftlichkeit, S. 13-26 in: Gehrke H.-J. (Hrsg.)(1994), Rechtskodifizierung und soziale Normen im interkulturellen Vergleich, Gunter Narr Verlag, Tübingen, 1994.

Kittredge R. I. (2003), Sublanguages and controlled languages, S. 430-447 in: The Oxford Handbook of Computational Linguistics, Mitkov R. (Hrsg.) (2003), Oxford University Press Oxford 2003.

Knorz G. (1997), Indexieren, Klassieren, Extrahieren, S. 120-140 in: Buder M., Rehfeld W., Seeger T., Strauch D. (Hrsgg.) (1997), Grundlagen der praktischen Information und Dokumentation, Saur Verlag München u.a., 4. Aufl. 1997.

Koch, P., Oesterreicher, W. (1994): Schriftlichkeit und Sprache, in: Hartmut Günther., Otto Ludwig (Hrsg.) (1994), Schrift und Schriftlichkeit. Ein interdisziplinäre Handbuch internationaler Forschung, 1. Halbband, Berlin/New York 1994.

Kurohashi, S., Nagao, M. (1998), Building a Japanese parsed corpus while improving the parsing system, S. 719-724 in: Rubio A., Gallardo N., Castro R., Tejada A. (Hrsgg.) (1998), First International Conference on Language Resources & Evaluation, Granada (Spanien) 1998.

Langer S. (2002), Grenzen der Sprachenidentifizierung, S. 99-106 in: Tagungsband KONVENS 2002, DFKI Saarbrücken 2002, http://konvens2002.dfki.de/cd/pdf/19V-langer.pdf: 27.9.2004.

Larkey L. S. und Croft W. B. (1996), Combining classifiers in text categorization, S. 289-297 in: Proceedings of SIGIR-96, 19. International Conference on Research and Development in Information Retrieval, ACM Press, New York, US, Zürich, CH, 1996.

Luckhardt H.-D., Harms I., Virtuelles Handbuch Informationswissenschaft, Universität des Saarlandes, http://www.is.uni-sb.de/studium/handbuch//exkurs.ind.php#intellind : 19.9.2004.

Lundquist, L. (1979), Teksttypbestemmelse af en lovtekst via en semantisk dybdestruktur, in: Linnarud, M., Svartvik, J. (Hrsgg.) (1979), Kommunikativ Kompetens och fackspråk, SyMPOSIUM Södertälje 1978, ASLA Uppsala 1979.

Malinowski, B. (1923), The Problem of Meaning in Primitive Languages, Supplement I auf S. 296-336 in: Ogden C. K., Richards I. A. (Hrsgg.), The Meaning of Meaning, 8. Auflage Harcourt Brace & World New York 1946.

Mann W., Thompson, S. (1988), Rhetorical Structure Theory: toward a functional theory of text organization, S. 243-281 in: Text 8 (1988): 3.

Manning C. D., Schütze H. (1999), Foundations of Statistical Natural Language Processing, MIT Press Cambridge (USA) und London 1999.

Matthiessen C., Bateman J. A. (1991), Text generation an Systemic-Functional Linguistics - Experiences from English and Japanese, Pinter Publishers London 1991.

Mayer F. (1996), The representation of inconsistent relationships in termbanks, S. 225-232 in: Galinski C., Schmitz K.-D. (Hrsgg.) (1996), Terminology and Knowledge Engineering Conference 1996 (TKE ’96), Indeks Frankfurt a. M. 1996.

Mayer F. (1998), Eintragsmodelle für terminologische Datenbanken: ein Beitrag zur übersetzungsorientierten Terminographie, Forum für Fachsprachen-Forschung Bd. 44, Narr Tübingen 1998.

Mayer F. (2000), Terminographie im Recht: Probleme und Grenzen der Bozner Methode, S. 295-306 in Veronesi, D., Rechtslinguistik des Deutschen und Italienischen, Unipress Padova 2000.

Maynard D., Ananiadou S. (2001), Term extraction using a similarity-based approach, S. 53-89 in: Bourigault D., Jacquemin C., L’Homme M. C. (Hrsgg.) (2001), Recent Advances in Computional Terminology, Natural Language Processing Band 2, John Benjamins Publishing Company Amsterdam/ Philadelphia 2001,
http://citeseer.nj.nec.com/maynard99term.html : 5.10.2004.

McEnery T. (2003), Corpus Linguistics, S. 448-463 in: Mitkov R. (Hrsg.) (2003): The Oxford Handbook of Computational Linguistics, Oxford University Press Oxford 2003.

Mel’čuk I. (2001), Communicative Organization in Natural Language: The Semantic - Communicative Structure of Sentences, Studies in Language Companion Series 57, John Benjamins Publishing Company Amsterdam 2001.

Melby A. K., Wright S. A. (1999), Leveraging terminological data for use in conjunction with lexicographical resources, S. 544-569 in: Proceedings of the Fifth International Congress on Terminology and Knowledge Engineering Conference Innsbruck (TKE ’99), TermNet Innsbruck 1999,
http://www.ttt.org/TKE-99.pdf :25.3.2004.

Merkel M., Mikael A. (2000), Knowledge-lite extraction of multi-word units with language filters and entropy thresholds, S. 737–746 in: Proceedings of the 6th International Conference on Content-Based Multimedia Information Access (RIAO) 2000, Vol. 1, Collège de France Paris 2000,
http://citeseer.nj.nec.com/merkel00knowledgelite.html : 4.10.2004.

Merkel M., Nilsson B., Ahrenberg L. (1994), A phrase-retrieval system based on recurrence, S. 43-56 in: Proceedings of the Second Annual Workshop on Very Large Corpora (WVLC-2), Kyoto 1994,
http://www.ida.liu.se/~magme/publications/kyoto--94.pdf : 13.9.2004.

Nakagawa H. (2001), Experimental evaluation of ranking and selection methods in term extraction, S. 303-325 in: Bourigault D., Jacquemin C., L’Homme M. C. (Hrsgg.) (2001), Recent Advances in Computional Terminology, Natural Language Processing Band 2, John Benjamins Publishing Company Amsterdam/ Philadelphia 2001,

http://www.r.dl.itc.u-tokyo.ac.jp/~nakagawa/academic-res/termrec.pdf : 5.10.2004.

NISO Standard Z39.85-2001 (September 2001): http://www.niso.org/standards/resources/Z39-85.pdf

Nohr H. (2001), Automatische Indexierung, Einführung in betriebliche Verfahren, System und Anwendungen, Materialien zur Information und Dokumentation, Verlag für Berlin-Brandenburg, Potsdam, 2001.

Noy N. F., Sintek M., Decker S., Crubezy M., Fergerson R. W., Musen M. A. (2001), Creating Semantic Web Contents with Protégé-2000., S. 60-71 in: IEEE Intelligent Systems 16(2):60-71, 2001,
http://www-smi.stanford.edu/pubs/SMI_Reports/SMI-2001-0872.pdf : 5.10.2004.

Ogawa Y., Matsuda T. (1997), Overlapping statistical word indexing: a new indexing method for Japanese text, S. 226 – 234 in: Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval, Philadelphia (USA) 1997, ACM Press New York 1997, http://portal.acm.org/ft_gateway.cfm?id=258576&type=pdf.

Ott S. (2003), Linking und Framing: Ein Überblick über die Entwicklung im Jahre 2002, JurPC Web-Dok. 14/2003, http://www.jurpc.de/aufsatz/20030014.htm: 23.9.2003.

Ozawa T., Yamamoto M., Umemura K., Church K. W. (1999), Japanese word segmentation using similarity measure for IR, S. 89-96 in: Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition (NTCIR Workshop 1), NACSIS (National Center for Science Information Systems) (Hrsg.) Tokyo 1999.
http://research.nii.ac.jp/~ntcadm/workshop/OnlineProceedings/023-IR-Ozawa.pdf.

Paijmans H. (1997), Gravity Wells of Meaning: detecting Information-Rich passages in Scientific Texts, S. 520-536 in: Journal of Documentation 53(5), 1997,
http://pi0959.kub.nl/Paai/Onderw/V-I/Content/grav_wells.html : 5.10.2004.

Patzold S. (2000), Konflikte im Kloster, Matthiessen Husum 2000.

Pegoraro L., Reposo A. (1993), Le fonti del diritto negli ordinamenti contemporanei, Monduzzi Editore Bologna 1993.

Picht H., Schmitz K.-D. (Hrsgg.) (2001), Terminologie und Wissensordnung, TermNet Wien 2001.

Pirkola A., Keskustalo H., Leppänen E., Känsälä H., Järvelin K. (2002), Targeted s-gram matching: a novel n-gram matching technique for cross-and monolingual word form variants, in: Information Research 7 (2002): 2.

Quasthoff U., Biemann C., Wolff C. (2002), Named Entity Learning and Verification: Expectation Maximization in Large Corpora, S. 8-14 in: Roth D., Bosch A. van den (Hrsgg.) (2002), Proceedings of the 6th Workshop on Computational Language Learning (CoNLL), 2002 Taipei.

Rasmussen, K. W; Engberg, J. (1999), Genre Analysis of Legal Discourse, S. 113-132 in: Hermes - Journal of Linguistics 1999, 22.

Roli F., Giacinto G. (2002), Design Of Multiple Classifier Systems, S. 199-226 in: Bunke H., Kandel A. (Hrsgg.) (2002), Hybrid Methods in Pattern Recognition, World Scientific Publishing Co. Singapore 2002, http://citeseer.ist.psu.edu/552125.html : 26.9.2004.

Runte M., Beißwenger M., Storrer A. (2003), Modellierung eines Terminologienetzes für das automatische Linking auf der Grundlage von WordNet, S. 95-104 in: Kunze C., Lemnitzer L., Wagner A. (Hrsgg.) (2003), Anwendungen des deutschen Wortnetzes in Theorie und Praxis, GermaNet Workshop Proceedings, Tübingen 2003,
http://www.sfs.uni-tuebingen.de/ lsd/GermaNet-Workshop/TermNet-LDV.pdf : 5.10.2004.

Sager J. C. (1990), A practical Course in Terminology Processing, John Benjamins Publishing Company Amsterdam 1990.

Schlohmann A. (1906), in: Illustrierte Technische Wörterbücher, zit. nach S. 366-367 in Picht a.a.O.

Schmidt G. (2002), XBRL ist... das, was die Finanzwelt braucht!, in: Consultant - Steuern Wirtschaft Finanzen, 05/2002, Max Schimmel Verlag Würzburg 2002.

Schmidt-Wigger A. (1998), Building consistent terminologies, Poster in: Proceedings of the First Workshop on Computational Terminology (COMPUTERM'98) at the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics (COLING-ACL'98), ACL University of Montreal (Hrsgg.) Montreal 1998,
http://www.iai.uni-sb.de/docs/term2.pdf : 1.10.2004.

Schmitz K.-D. (2002), Towards a uniform environment for representing terminologies within ISO, Präsentation auf der Terminology and Knowledge Engineering (TKE) Konferenz Nancy 2002,
http://tke2002.loria.fr/Doc/workshops/ws2/ws2_kds.ppt :18.3.2004.

Shannon, C. E. (1998), The Mathematical Theory of Communication, University of Illinois Press Urbana/ Chicago 1998.

Shapiro S.C. (1992), Artificial Intelligence, S. 54-57 in: S. C. Shapiro, (Hrsg.) (1992), Encyclopedia of Artificial Intelligence, 2. Aufl. John Wiley & Sons Inc. New York 1992.

Sieber P. (1998), Parlando in Texten. Zur Veränderung kommunikativer Grundmuster in der Schriftlichkeit, S. 182-188 in: Henn H., Sitta H., Wiegand H.E. (Hrsg.) (1998), Reihe Germanistische Linguistik 191, Max Niemeyer Verlag Tübingen 1998.

Silber K.F., McCoy H.G. (2000), Efficient Text Summarization Using Lexical Chains, S. 252-255 in: Proceedings of the 5th international conference on Intelligent user interfaces, New Orleans 2000, http://web.media.mit.edu/~lieber/IUI/Silber/Silber.pdf: 5.10.2004.

Smadja F. (1993), Retrieving collocations from text: Xtract, S. 142-176 in: Computational Linguistics 19 (1),
http://acl.ldc.upenn.edu/J/J93/J93-1007.pdf: 4.10.2004.

Soininen P., Voutilainen A., Tapanainen P. (1999), An experiment in automatic term extraction, S. 234–240 in: Sandrini P. (Hrsg.) (1999), Proceedings of the 5th International Congress on Terminology and Knowledge Engineering (TKE) 1999, TermNet Innsbruck 1999.

Somers H. (2003), Machine Translation: Latest Developments, S. 512-528 in: Mitkov R. (Hrsg.) (2003), The Oxford Handbook of computational Linguistics, Oxford University Press Oxford 2003.

Streiter O. (2001), Corpus-based parsing and treebank development, S. 115-120 in: ICCPOL 2001, 19th International Conference on Computer Processing of Oriental Languages, Seoul (Korea) 2001.

Streiter O., De Luca E. W. (2003), Example-based NLP for Minority Languages: Tasks, Resources and Tools, S. 233-242 in: Proceedings of the Conférence sur le Traitement Automatique des Langues Naturelles (TALN), Association pour le Traitement Automatique des Langues (ATALA) Batz-sur-Mer 2003,
http://dev.eurac.edu:8080/taln/accepted/streiter_de_luca.pdf : 5.10.2004.

Streiter O., Hsueh P. (2000), A case-study on example-based parsing, in: Tagungsband der International Conference on Chinese Language Computing (ICCLC) Chicago, 2000.
http://citeseer.nj.nec.com/streiter00case.html : 23.12.2003.

Streiter O., Stuflesser M., Ties I. (2004), CLE, an aligned Trilingual Ladin-Italian-German Corpus. Corpus Design and Interface, S. 84-87 in: Carson-Berndsen J. (2004), Proceedings of the SALTMIL Workshop at LREC 2004, First Steps in Language Documentation for Minority Languages – Computational Linguistic Tools for Morphology, Lexicon and Corpus Compilation, European Language Resources Association (ELRA) Paris 2004,
http://dev.eurac.edu:8080/autoren/publs/lrec9q.pdf : 30.3.2004.

Streiter O., Voltmer L. (2003), A Model for Dynamic Term Presentation, S. 201-204 in: Tagungsband der TIA-2003 Konferenz, LIIA – ENSAIS, Université Marc Bloch (Hrsgg.) Strasbourg 2003, http://dev.eurac.edu:8080/autoren/publs/ModDynTerm.pdf.gz: 13.9.2004.

Streiter O., Voltmer L. (2003), Text Classification for Corpus-Based Legal Terminology, S. 253-260 in: Vrabie, G., Turi, J.G. (Hrsg.) (2003), La théorie et la pratique des politiques linguistiques dans le monde, Tagungsband der 8. Internationalen Konferenz der Académie Internationale de Droit Linguistique 2002, Editura CUGETAREA Iaşi (Rumänien) 2003.

Streiter O., Zielinski D., Ties I., Voltmer L. (2002), Example-based Term Extraction for Minority Languages: A case-study on Ladin, in: Tagungsband von Soziolinguistica Y Language Planning, SPELL St. Ulrich (Italien) noch in Vorbereitung,
http://dev.eurac.edu:8080/autoren/publs/termex5.pdf : 5.10.2004.

Streiter O., Zielinski D., Ties I., Voltmer L. (2003), Term Extraction for Ladin: An Example-based Approach, S. 275-284 in: Tagungsband der Konferenz zum Traitement Automatique des Langues Naturels (TALN) 2003, Batz-sur-Mer, 2003,
http://dev.eurac.edu:8080/autoren/publs/taln_minority_streiter_et_all.pdf : 5.10.2004.

Teich E., Fankhauser P. (2004), WordNet for Lexical Cohesion Analysis, S. 326-331 in: Sojka P., Pala K., Smrž P., Fellbaum C., Vossen P. (Hrsgg.) (2004), Proceedings of the Second International WordNet Conference (GWC 2004), Masaryk Universität Brno 2003.

Umstätter W. (1992), Nutzen der Indexierung bei Online-Datenbanken, S. 403-420 in: 14. Online-Tagung 1992 Frankfurt am Main, DGD-Schrift (OLBG-13) 2/92,
http://www.ib.hu-berlin.de/~wumsta/pub65.html: 17.11.2003.

Van Rijsbergen C. J. (1983), Information Retrieval, Butterworths London 1983.

Voß J. (2004), Begriffssysteme – Ein Vergleich verschiedener Arten von Begriffssystemen und Entwurf des integrierenden Thema Daten¬modells, Studienarbeit im Diplomstudiengang Informatik, HU Berlin, Version 1.1d,
http://www.nichtich.de/epub/begriffssysteme03/begriffssysteme.pdf : 8.3.2004.

Vossen P. (2003), Ontologies, S. 464-482 in: The Oxford Handbook of Computational Linguistics, Mitkov R. (Hrsg.) (2003), Oxford University Press Oxford 2003.

Wendt H. (1987), Fischer Lexikon: Sprachen, Fischer Taschenbuchverlag, Frankfurt am Main, 1987.

Widdows D., Peters S., Cederberg S., Chan C.-K., Steffen D., Buitelaar P. (2003), Unsupervised Monolingual and Bilingual Word-Sense - Disambiguation of Medical Documents using UMLS, S. 9-16 in: Natural Language Processing in Biomedicine, ACL 2003 workshop proceedings, Association for Computational Linguistics Sapporo 2003.
http://citeseer.ist.psu.edu/584877.html : 18.6.2004.

Zielinski D. (2002), Computergestützte Termextraktion aus technischen Texten, Diplomarbeit Universität des Saarlands, Saarbrücken 2002,
http://www.iai.uni-sb.de/~mt-dept/texte/zielinski.pdf: 4.10.2004

.

# Index

**—X—**

# Lebenslauf von Leonhard Anton Georg **Voltmer**

Deutscher, geb. am 20. März 1971 in München
Dreiheiligengasse 8, 39100 Bozen, Italien
Tel. privat: +39 3283830254, Fax geschäftlich: +39 0471 055-299
Email: LVoltmer@web.de

## Berufliches

| | | | |
|---|---|---|---|
| **Staatskundelehrer** | Gymnasium Rosmini in Trient | Italien | seit Okt. 2003, Teilzeit |
| **Wiss. Mitarbeiter** | Europäische Akademie Bozen | Italien | seit Juli 2001 |
| **Rechtsanwalt** | Selbständig | Deutschland | seit 2000, Teilzeit |
| **Studienreiseleiter** | Studiosus Reisen | Europa | 2000 bis2004 Saisonarbeit |
| **Rechtsreferendar** | in Bayern und an der Deutschen Botschaft Madrid | Deutschland und Spanien | 1997 bis 1999 |
| **Praktikant** | Landgericht Lund | Schweden | März 1998 |
| **Praktikant** | Deutscher Bundestag | Deutschland | September 1995 |
| **Wiss. Mitarbeiter** | LMU München | Deutschland | 1996 und 1998 |

## Bildung

| | | | |
|---|---|---|---|
| **Promotionsstudium** zum Dr. Phil. Romanistik, Computerlinguistik, Jura | LMU München | Deutschland | 2002-2005 |
| **Romanistikstudium** (Italienisch und Portugiesisch) | Universität Salzburg | Österreich | 2000-2001 |
| **Studienreiseleiterausbildung** | Studiosus München | Deutschland | 03/2000 |
| **2. jur. Staatsexamen**: 5,72 Punkte | Justizministerium Bayern | Deutschland | 12/1999 |
| **Master in Rechtstheorie** LL.M. cum laude | Europ. Akademie f. Rechts-theorie Brüssel und Univ. Lund | Belgien, Schweden | 08/1997 |
| **Doppelstudium** Jura - Romanistik<br>**1. jur. Staatsexamen**: 7,45 Punkte | LMU München<br>Stipendium des DFHK | Deutschland | 1991-93<br>1994-98 |
| **Licence en Droit** | Univ. Paris II Assas-Panthéon | Frankreich | 06/1994 |
| **Abitur** Durchschnitt 1,9 | Gymnasium Pullach | Deutschland | 1990 |

**Englisch**, **Französisch**, **Spanisch**, **Schwedisch** und **Italienisch** verhandlungssicher.

Weitere Kenntnisse in und **Portugiesisch** und **Niederländisch**.

## Sprachwissenschaftliche Kompetenzen

- Erfahrener Übersetzer, Dolmetscher und Terminologe.
- Spezialist für juristische Datenbanken, Markup-Sprachen und Data-Mining.
- Mitarbeit an Konzeption und Realisierung der Sprachlernplattform Gymn@zilla.
- Mitarbeit an Konzeption und Aufbau des Rechtsinformationssystems BISTRO.

Lebenslauf Online: http://www.eurac.edu/About/Collaborators/LVoltmer/Lebenslauf.htm

Veröffentlichungen: http://www.eurac.edu/About/Collaborators/LVoltmer/Veröffentlichungen.htm