LDV-FORUM

Zeitschrift für Computerlinguistik und Sprachtechnologie
GLDV-Journal for Computational Linguistics and Language Technology

LDV-Forum 15 (1998) 1
Zeitschrift für Computerlinguistik und Sprachtechnologie
GLDV-Journal for Computational Linguistics and Language Technology
Offizielles Organ der GLDV

Herausgeber
Prof. Dr. Gerhard Knorz; Gesellschaft für Linguistische Datenverarbeitung
Anschrift: Fachhochschule Darmstadt, Fachbereich Information und Dokumentation, Haardtring 100, D-64289 Darmstadt; Tel: (0 6151) 16-8499; Fax: (0 6151) 16-8980; e-mail: knorz@iud.fh-darmstadt.de

Redaktion
Gerhard Knorz

Wissenschaftlicher Beirat
Prof. Dr. W. Hoeppner (hoeppner@unidui.uni-duisburg.de);
Prof. Dr. Gerhard Knorz;
Prof. Dr. Winfried Lenders (lenders@uni-bonn.de);
Prof. Dr. Ulrich Schmitz (ulrich.schmitz@uni-essen.de)

Erscheinungsweise:
2 Hefte im Jahr, halbjährlich zum 31. Mai und >>>

Editorial

Das LDV-Forum scheint in Format und Aufmachung auf den ersten Blick unverändert, aber das täuscht! Zwar ist der Titel LDV-Forum genauso geblieben wie auch der Name der GLDV, aber die erklärenden Untertitel sind nunmehr andere. Jüngstes (mir bekanntes) Vorbild ist die GMD, die inzwischen unter Beibehaltung ihres Akronyms auch keine „Gesellschaft für Mathematik und Datenverarbeitung“ mehr ist, sondern „GMD – Forschungszentrum Informationstechnik GmbH/National Research Center for Information Technology“. Ein viel älteres Beispiel, wenn auch nicht ganz so auffällig, liefert das LDV-Forum selbst: Der Name stammt nämlich noch aus einer Zeit, in der die heutige GLDV noch den Namen LDV-Fittings trug. Jüngere Mitglieder werden sich schon gewundert habe, wieso die Zeitschrift eigentlich nicht GLDV-Forum heißt. Alles klar nun? Man muß es nicht so verunglückt halten wie eine große Kreditkartenfirma, die empfiehlt, mit dem eigenen guten Namen zu bezahlen, aber man wechselt nicht leichtfertig einen guten Namen!

Nun bedarf es bei Änderungen dieser Art einer überzeugenden Begründung, die allerdings in vorliegendem Fall nicht schwerfällt. Ich für meine Person will es damit bewenden lassen, daß ich die die Aktivitäten, die sich u.a. in diesem Ergebnis der Namensdiskussion niederschlagen, als Mitglied des

Beirates der GLDV mit voranbringe und sehr begrüße. Und ansonsten verweise ich auf den nachfolgen-den programmatischen Beitrag des neuen Vorstands der GLDV, der diesmal das eigentliche Editorial zu dieser Ausgabe darstellen soll. In ungewohnter, aber vielleicht wohltuender (?) Kürze

G.K.

PS: Der Buchstabe dieses Heftes, das „V" (sprich: „Vogel-Vau") drängt sich diesmal nicht unmittelbar auf. Das kann allerdings nicht darüber hinwegtäuschen, daß das neue Jahrtausend mit dem Wendejahr 2 000 als eine seiner Herausforderungen die Aufgabe bereithält, ein Titel-Konzept für die Zeit nach dem „Z" zu finden.

>>> 31. Oktober. Preprints und redaktionelle Planungen sind laufend und aktuell unter dem URL http://www.iud.fh-darmstadt.de/iud/wwmeth/publ/ldvforum/menu1.htm einsehbar.

Bezugsbedingungen: Für Mitglieder der GLDV ist der Bezugspreis des LDV-Forum im Jahresbeitrag mit eingeschlossen. Jahresabonnements können zum Preis von DM 40,- (incl. Versand), Einzelexemplare zum Preis von DM 20,- (zuzügl. Versandkosten) bestellt werden: LDV-Forum, c/o IKS, Poppelsdorfer Allee 47, 53 115 Bonn.

Fachbeiträge: Unaufgefordert eingesandte Fachbeiträge werden vor Veröffentlichung von mindestens 2 ReferentInnen begutachtet. Manuskripte sollten deshalb möglichst frühzeitig eingereicht werden und bei Annahme zur Veröffentlichung in jedem Fall elektronisch und zusätzlich auf Papier übermittelt werden. Artikel werden am besten in Microsoft Word™ für Windows oder Word Perfect™ für Windows erstellt. Eine Dokumentvorlage für Word™ für Windows, in der die wichtigsten Styles enthalten sind, kann unter dem URL ftp://www.iud.fh-darmstadt.de/iud/wwmeth/publ/ldvforum/ldvforum.dot heruntergeladen werden.

Rubriken: Die namentlich gezeichneten Beiträge geben ausschließlich die Meinung der AutorInnen wieder. Einreichungen sind – wie bei Fachbeiträgen – an den Herausgeber zu übermitteln.

Redaktionsschluß: Für Heft 15 (1998) 2: 15. August 1998.

Druck und Vertrieb: IKS, Poppelsdorfer Allee 47, 53 115 Bonn. **Herstellung:** Kurt Thomas, IKP, Universität Bonn. e-mail: thomas@uni-bonn.de . **Auflage:** 400 Exemplare. **Anzeigen:** Preisliste und Informationen: IKS e. V., Poppelsdorfer Allee 47, D-53 115 Bonn, Tel.: (0 228) 73 5621, Fax: (0 228) 73 5639, e-mail: iks@uni-bonn.de .
Bankverbindung: IKS e. V.: PGA Köln (BLZ 370 100 50), Konto 385647-505.

Die vorliegende Ausgabe wurde u. a. in Divona gesetzt. Dieser **Font** wurde uns freundlicherweise von der Firma Scriptorium (http://www.ragnarokpress.com/scriptorium) zur Verfügung gestellt.

GLDV-Anschrift: Prof. Dr. R. Hausser, Universität Erlangen-Nürnberg, Abteilung für Computerlinguistik, Bismarckstraße 12, D-91 054 Erlangen; e-mail: rrh@linguistik.uni-erlangen.de .

GLDV – Gesellschaft für linguistische Datenverarbeitung – Society for Computational Linguistics and Language Technology

Roland Hausser, 1. Vorsitzender
Universität Erlangen-Nürnberg
e-mail: rrh@linguistik.uni-erlangen.de

Im Frühjahr 1997 wurde turnusmäßig ein neuer Vorstand und Beirat gewählt. Deshalb möchte ich zunächst dem alten Vorstand und Beirat für ihre Arbeit danken und anschließend darstellen, welchen Kurs die GLDV in der näheren Zukunft steuern soll.

Mein besonderer Dank gilt dem bisherigen 1. Vorsitzenden, Prof. Dr. Winfried Lenders. Als erfahrener Organisator – ich erinnere z. B. an die CoLing 1986 und die GLDV-Herbstschule zu ‚modernen Methoden der Lexikologie' 1996 – hat er unsere Gesellschaft mit ruhiger Hand von 1993 – 97 geführt. Er wird auch die Konvens IV (5. – 7. Oktober 1998 in Bonn) ausrichten, unter deren Dach die nächste GLDV-Mitgliederversammlung stattfindet. Ich freue mich, daß Winfried Lenders der GLDV als gewähltes Mitglied des Beirats weiterhin aktiv verbunden bleibt. Zugleich begrüße ich Herrn Dr. Hans Klaus (GMD Darmstadt, Projektträger Fachinformation) und Herrn Jochen Leidner (Student der Computerlinguistik) als kooptierte Mitglieder des Beirats.

Was den neuen Kurs der GLDV angeht, waren sich der neue Vorstand und Beirat in folgenden Punkten einig. Erstens bedarf das Image dieser ältesten computerlinguistischen Gesellschaft im deutschsprachigen Raum eines gründlichen Face-Lifts. Zweitens muß die Zukunft der Gesellschaft durch den Zugewinn neuer studentischer Mitglieder langfristig gesichert werden. Diese beiden Ziele stehen in einem direkten Zusammenhang und sind durch eine verbesserte Image- und Informationspflege unserer Gesellschaft im elektronischen Medium erreichbar. Als professioneller Fachverband für Computerlinguistik und Sprachtechnologie ist die GLDV geradezu prädestiniert, ihre Aufgabe als Vermittler zwischen Studenten, akademischen Instituten, der einschlägigen Industrie und der weiteren Öffentlichkeit mit elektronischen Kommunikationsmitteln zu erfüllen. Zugleich kann sie auf diese Weise ihre praktische Expertise unter Beweis stellen und für ihr Fachgebiet in einer medien-gerechten, allgemeinverständlichen Form werben.

Deshalb wurde zunächst (i) die GLDV-Homepage überarbeitet, die bereits bei vielen ‚Besuchern' großen Anklang gefunden hat. In diesem Zusammenhang darf ich auf den Aufsatz von Herrn Dr. Bernhard Schröder, dem neuen Informations-

referenten der GLDV (b.schroeder@uni-bonn.de), und Herrn Hans-Christian Schmitz in diesem Heft verweisen, der Design und Struktur der aktuellen GLDV-Webseiten erläutert.

Ein weiterer Teilaspekt (ii) von GLDV-Aktivitäten, bei dem die Elektronik eine besonders effiziente und flexible Durchführung ermöglicht, ist die Information und Befragung der Mitglieder. Diese Vorgänge sind bisher auf *hard copy* (Mitteilungen im LDV-Forum, Briefwahlen) oder persönliche Teilnahme an den jährlichen Mitgliederversammlungen beschränkt gewesen. Als Ergänzung wäre es jedoch praktisch, e-mail zu verwenden. Bisher waren aber nur die e-mail-Adressen von weniger als 100 Mitgliedern bekannt. Deshalb wurden eine Telephonaktion gestartet, um die übrigen e-mail-Adressen mit möglichst geringem bürokratischen Aufwand für die Befragten zu ermitteln. Die vervollständigte e-mail-Liste wird schon jetzt zur Information (z. B. über Aktivitäten der Arbeitskreise) verwendet. Ob die Mehrheit der Mitglieder damit einverstanden ist, daß in Zukunft auch Befragungen (Abstimmung zu Themen der Mitgliederversammlung, Wahlen) durch e-mail stattfinden, wird auf der nächsten Mitgliederversammlung als Tagesordnungspunkt diskutiert und zur Abstimmung gebracht werden. In diesem Zusammenhang steht der Beitrag des neuen Schriftführers, Herrn Dr. Christian Wolff (wolff@informatik.uni-leipzig.de), in diesem Heft, der die technischen Aspekte einer elektronischen Wahl behandelt. Ein anderer Bereich (iii), in dem Handlungsbedarf besteht, ist eine Neufassung des GLDV-Studienführers und -Ausbildungsprofils, die zuletzt im Jahre 1991 überarbeitet bzw. herausgebracht wurden. Dieses aufwendige Unternehmen erfordert Anfragen bei den CL-Instituten, um so die Angaben auf den neusten Stand zu bringen. Die Darstellung des Studienführers im Netz soll dadurch vereinfacht und verbessert werden, daß speziellere Informationen wie z. B. das aktuelle Studienangebot über Links direkt von den Homepages der einzelnen Institute zur Verfügung gestellt und aktualisiert werden. Weitere Veränderungen und Aktivitäten betreffen (iv) die Tätigkeit der Arbeitskreise, (v) den Aufbau einer möglichst vollständigen bundesweiten Liste mit den Themen von CL-Abschlußarbeiten ab 1.1.1998 und den anschließenden Berufen der Kandidaten, (vi) ein elektronisches Verfahren zur Aufnahme neuer Mitglieder, (vii) die Einrichtung von GLDV-bboards, (viii) eine verstärkte Kooperation mit der einschlägigen Industrie (auch für Praktikumsplätze) und (ix) eine neue inhaltliche und optische Gestaltung des Faltblatts. Mitglieder von Vorstand und Beirat werden bei Gelegenheit auf diese und eine Reihe weiterer Punkte zurückkommen.

Als nächste größere GLDV-Veranstaltung wurde beschlossen, vom 28. September bis zum 2. Oktober 1998 eine GLDV-*Herbstschule* zu dem Thema ***Web-Linguistik:*** *Sprachtechnologie für das Internetz* abzuhalten, und zwar an der Abteilung Computerlinguistik der Universität Erlangen-Nürnberg. Es werden die folgenden Kurse angeboten:

- *Hypertext und Textdatenbanken im World Wide Web*
  Angelika Storrer, IDS Mannheim & Roman Schneider, Oracle
- *Informationsmodellierung in XML und SGML*
  Henning Lobin, Universität Bielefeld
- *Web-basierte maschinelle Übersetzung*
  Uta Seewald, Universität Hannover & Rita Nübel,
  IAI Saarbrücken
- *Unicode*
  Carl-Martin Bunz, Universität Saarbrücken &
  Koaunghi Un, Universität Tübingen
- *Multimedia*
  Jürgen Handke, Universität Marburg
- *Text-Mining-Technology*
  Sebastian Göser, IBM Stuttgart

Um neue studentische GLDV-Mitglieder zu werben, erhalten Teilnehmer an dieser Herbstschule ein ‚Schnupperangebot' (erheblich reduzierter Jahresbeitrag für das erste Jahr). Last – but not least – wurde vom Vorstand beschlossen, die Namensverwendung der GLDV und des LDV-Forum durch zusätzliche englische Untertitel zu ergänzen. Indem unsere Gesellschaft nun als *GLDV – Gesellschaft für linguistische Datenverarbeitung – Society for Computational Linguistics and Language Technology* und das Forum als *LDV-Forum, Zeitschrift für Computerlinguistik und Sprachtechnologie – Journal for Computational Linguistics and Language Technology* firmieren, soll ihre Thematik einem weiteren Publikum in einer zeitgemäßen Terminologie verständlich gemacht werden. Zugleich wird durch Beibehaltung des alten Namens der bürokratische Aufwand einer Namensänderung vermieden und die Tradition unserer Gesellschaft fortgesetzt.

Zum Schluß möchte ich, auch im Namen des Vorstands und Beirats der GLDV, allen Mitgliedern der Gesellschaft an dieser Stelle ausdrücklich für das Vertrauen danken, das sie uns durch ihre Wahl ausgesprochen haben. Wir werden uns nach Kräften bemühen, diesem gerecht zu werden.

Häufigkeitsverteilung deutscher Morpheme

Roland Hausser
Universität Erlangen-Nürnberg
Abteilung Computerlinguistik (CLUE)
rrh@linguistik.uni-erlangen.de

Abstract

Bisher bezogen sich Angaben zum Wortschatz einer Sprache meist auf Wortformen und basierten auf Korpora, die möglichst balanciert und repräsentativ sein sollten. Die vorliegende Untersuchung betrachtet neben der Verteilung der Wortformen auch die der Morpheme und Allomorphe, basierend auf einer regelgesteuerten automatischen Wortformerkennung (DMM). Die Morphemverteilung in einem klassischen Korpus wird mit der in domänenspezifischen Korpora verglichen.

1. Desiderata der Korpuskonstruktion

Bereits im Jahre 1897–98 präsentierte Wilhelm Kaeding die erste umfassende Häufigkeitsuntersuchung von Wortformen für das Deutsche, und zwar als statistische Grundlage zur Verbesserung der Stenographie.[1] Mit seiner kleinen Armee von ‚Zählern' untersuchte Kaeding fast 11 Millionen laufende Wortformen (= 20 Millionen Silben) mit 250 178 verschiedenen Wortformen (Types).

Damit war die von Kaeding verwendete Textmenge mehr als zehnmal so groß und die resultierende Menge der Types mehr als doppelt so groß wie die des computerbasierten Limas-Korpus von 1973. Im Gegensatz zu heutigen Korpora handelt es sich bei Kaedings Textsammlung allerdings nicht um ein streng synchrones Korpus, denn die von ihm ausgewählten Beispiele decken den Zeitraum von ca. 1750 bis 1890 ab.

Mit der Verfügbarkeit von Computern erhielten Untersuchungen dieser Art neue Impulse, wobei Kučera und Francis 1967 für amerikanisches Englisch mit dem Brown-Korpus[2] den Anfang machten. Das Brown-Korpus umfaßt 500 Texte mit 1 014 231 laufenden Wortformen (Tokens) und 50 406 verschiedenen Wortformen (Types).

1968 folgte das LOB-Korpus[3] als Pendant zum Brown-Korpus für britisches

[1]Siehe Meier 1964.

[2]Benannt nach der Brown University in Rhode Island, an der Francis lehrte.

Englisch, ebenfalls mit 500 Texten, ca. einer Million Tokens und 50 000 Types. Beide Korpora wurden aus Texten der folgenden 15 *Genres* zusammengestellt.

		Brown	LOB
A	Presse: Reportagen	44	44
B	Presse: Kommentare	27	27
C	Presse: Rezensionen	17	17
D	Religion	17	17
E	Handwerk, Handel und Freizeit	36	38
F	Trivialliteratur	48	44
G	Literatur, Biographien, Essay	75	77
H	Regierungsdokumente etc.	30	38
J	Geistes- und naturwissenschaftliche Schriften	80	80
K	Erzählungen allgemein	29	29
L	Kriminalromane	24	24
M	Science-fiction	6	6
N	Abenteuer- und Wildwestgeschichten	29	29
P	Liebesromane	29	29
R	Humor	9	9
Gesamt		500	500

1.1 Die 15 Genres des Brown- und des LOB-Korpus

Die Zahlen besagen, wieviele Texte aus dem jeweiligen Genre in das Korpus aufgenommen wurden – wobei leichte Differenzen zwischen dem Brown- und dem LOB-Korpus festzustellen sind.

Für den Bau des Brown-Korpus formulierten Kučera und Francis 1967, S. xviii, folgende Desiderata:

> **1.** Exakte Spezifikation der verwendeten Sprachtexte, so daß sich die Benutzer einen genauen Begriff von der Zusammensetzung des Materials machen können.
>
> **2.** Vollständige Synchronizität: Nur Texte aus einem einzigen Kalenderjahr werden verwendet.

[3]Das *Lancaster-Oslo/Bergen*-Korpus wurde unter der Leitung von Geoffrey N. Leech und Stig Johannson angelegt. Siehe Hofland & Johannson 1982.

3. Die verschiedenen Genres werden in einem vorgegebenen Größenverhältnis zueinander gefüllt, wobei die individuellen Textbeispiele nach dem Zufallsprinzip ausgewählt werden *(random sampling)*.

4. Formale Spezifikation der im Korpus enthaltenen Informationen und automatischer Zugriff auf sie.

5. Genaue und vollständige Beschreibung der elementaren statistischen Eigenschaften des Korpus und seiner Komponenten (Genres), mit der Möglichkeit, die Analyse auf Erweiterungen des Korpus auszudehnen.

1.2 Desiderata der Korpuskonstruktion

Diese Ansprüche werden umgesetzt mit den mathematischen Methoden der Statistik, also den Grundgleichungen der Stochastik, Verteilungen für unabhängige und abhängige Häufigkeiten, Normalisierung, Fehlerberechnung etc. Dabei wird versucht, eine theoretische Verteilung zu finden, der die empirische Verteilung entspricht, insbesondere bei beliebig wachsender Datenmenge (Konstanz der empirischen Verteilungsverhältnisse).

Die deutsche Entsprechung zum amerikanischen Brown-Korpus (1967) und dem britischen LOB-Korpus (1968) ist das Limas-Korpus (1973).[4] Wie seine englischen Pendants besteht es aus 500 Texten von jeweils ca. 2 000 laufenden Wortformen. Insgesamt enthält das Limas-Korpus 1 062 624 laufende Wortformen. Aufgrund der reicheren Morphologie des Deutschen ist die Zahl der Types mit 110 837 verschiedenen Wortformen jedoch mehr als doppelt so groß wie bei den englischen Korpora.

2 Auswahl der Genres

Die Auswahl der Genres und die Festlegung ihrer unterschiedlichen Größe haben das Ziel, ein Korpus möglichst *repräsentativ* für die Sprache seiner Zeit zu machen und die verschiedenen Genres möglichst *balanciert* zu vertreten.[5] Intuitiv sind diese Begriffe leicht verständlich. So ist z. B. der Jahrgang einer Tageszeitung repräsentativer für eine natürliche Sprache als die Summe der Telefonbücher oder die Kontoauszüge einer Bank. Und ein Korpus, das Texte aus den verschiedenen Genres in den Verhältnissen von 1.1 enthält, ist besser balanciert als eines, das nur aus Texten eines einzigen Genres besteht.

[4]Siehe Hess, K., J. Brustkern & W. Lenders 1983.

[5]Siehe Bergenholz 1989, Biber 1994, Oostdijk & de Haan (Hg.) 1994.

Dennoch ist es schwierig, die für ein Korpus gewählte Zusammensetzung als repräsentativ und balanciert zu *beweisen*. Oostdijk 1988 kritisiert z. B. an 1.1:

> *[...] 'genre' is not a well-defined concept. Thus genres that have been distinguished so far have been identified on a purely intuitive basis. No empirical evidence has been provided for any of the genre distinctions that have been made [...].*

Für ein wirklich repräsentatives und balanciertes Korpus ist letztlich erforderlich, daß man weiß, welche Genres wie oft in einem gegebenen Zeitraum von der Sprachgemeinschaft gesprochen, geschrieben, gehört und gelesen wurden. Da es praktisch unmöglich ist, das Verhältnis zwischen Produktion und Rezeption sowohl gesprochener als auch geschriebener Sprache in sämtlichen Genres realistisch zu quantifizieren, ist der Bau repräsentativer und balancierter Korpora naturgemäß mehr das Ergebnis einer Kunst als einer Wissenschaft. Es beruht weitgehend auf allgemeinen ‚common sense'-Überlegungen und hängt zudem von dem intendierten Zweck des Korpus ab.

Inzwischen werden Korpora wie das Brown-, LOB- und Limas-Korpus im Umfang von jeweils 1 Million laufender Wortformen als wesentlich zu klein für die Erstellung aussagekräftiger Statistiken im Bereich der natürlichen Sprachen angesehen. Deshalb wurde für das britische Englisch das British National Corpus (BNC) mit 100 Millionen laufenden Wortformen zusammengestellt. Davon sind 89,7 Millionen aus dem Bereich geschriebener Sprache und 10,34 Millionen aus dem Bereich der gesprochener Sprache. Der Bereich der geschriebenen Sprache umfaßt 659 270 Types[6].

3 Auswertung von Korpora

Der Wert eines Korpus liegt nicht in dem Inhalt seiner Texte, sondern in seiner Eigenschaft als reale Stichprobe einer natürlichen Sprache. Je repräsentativer und balancierter diese Stichprobe ist, desto wertvoller ist das Korpus – zum Beispiel für eine realistische Berechnung der statistischen *Häufigkeitsverteilung* der Wortformen.

Auf der elementarsten Ebene wird diese statistische Auswertung als eine *frequenzbasierte* und eine *alphabetische* Wortliste dargestellt (als Beispiele siehe 4.1 und 4.3). In der frequenzbasierten Wortformenliste werden die Types in der

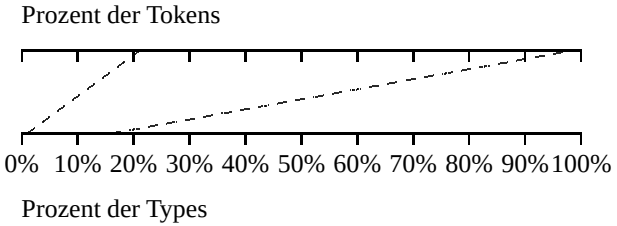
[6]Diese Zahl bezieht sich auf die reinen Oberflächen der Wortformen. Die Autoren des BNC verwenden dagegen Zahlen, die auf getaggten Wortformen beruhen. Nach letzterer Zählweise enthält das BNC 921 073 Types.

Reihenfolge ihrer (Token-) Häufigkeit aufgelistet. Der Platz einer Wortform in dieser Reihenfolge wird der *Rang* der Wortform genannt.

Am Anfang der Frequenzliste des BNC steht zum Beispiel die Wortform the, die mit 5 776 399 Vorkommen 6,4 % der laufenden Wortformen (Tokens) ausmacht. Am unteren Ende der Frequenzliste stehen die Wortformen, die jeweils nur ein einziges Mal im Korpus vorkommen. Von diesen sogenannten *Hapax Legomena*[7] gibt es 348 145 – was 52,8 % der Types im BNC ausmacht.

Wenn wir die ersten neun Ränge betrachten, so umfassen sie 0,001368 % der Types, decken aber 21,895 % der laufenden Wortformen im BNC ab. Am anderen Ende der Skala liegen die Ränge, deren Wortformen nur ein- bis neunmal im BNC vorkommen. Sie umfassen 83,6 % der Types, entsprechen aber zusammen nur 1,2 % der laufenden Wortformen.

Mit anderen Worten, 16,4 % der Types im BNC genügen, um 98,8 % der laufenden Wortformen zu erfassen. Für das verbleibende 1,2 % der laufenden Wortformen werden 83,6 % der Types im BNC benötigt. Das entspricht dem Intervall zwischen Rang 659 270 und 108 155, also insgesamt 551 115 Types. Diese Korrelation von Type- und Token-Häufigkeit findet sich ganz allgemein in ausreichend großen Korpora natürlicher Sprachen. Sie ist in 3.1 noch einmal graphisch dargestellt.



3.1 Korrelation von Type und Token-Häufigkeit

Die Tatsache, daß 0,001 % der Types fast 22 % der laufenden Wortformen in einem Korpus abdecken, und 16 % der Types über 98 % der laufenden Wortformen, wird manchmal dahingehend mißverstanden, daß kleine Lexika, wie z. B. die heutiger Spracherkennungssysteme mit nur 1 000 Wortformen, für die meisten praktischen Belange vollkommen genügen würden. Dies ist jedoch insofern ein

[7] Aus dem Griechischen, „einmal gesagte".

schwerwiegender Irrtum, als die *semantische Signifikanz*[8] mit abnehmender Frequenz einer Wortform steigt.

So nützt es dem Benutzer herzlich wenig, wenn das System zwar der, ist, mit und zu, nicht aber signifikante Hapax Legomena wie z. B. Abbremsung, Babyflaschen oder Campingplatz versteht,[9] weil es in seinem Lexikon keinen Eintrag dafür gibt. Für das BNC gilt entsprechendes: Unter seinen Hapax Legomena finden sich z. B. audiophile, butternut, customhouse, dustheap, um nur wenige Beispiele zu nennen, die alle in einem traditionellen Lexikon wie z. B. dem Webster's New Collegiate Dictionary aufgeführt sind und dort lexikalisch beschrieben werden.

Darüber hinaus gibt es viele Wörter im Webster's, die im BNC kein einziges Mal belegt sind, z. B. *aspheric*, *bipropellant*, *dynamotor* – trotz seiner Größe und des Bemühens um ein repräsentatives, balanciertes Korpus. Somit kann die Type-Liste eines großen Korpus zwar helfen, ein traditionelles Lexikon zu ergänzen. Es ist jedoch nicht zu erwarten, daß sich ein großes Korpus als ebenso vollständig wie oder vollständiger als ein traditionelles Lexikon erweist.

4 Statistisches Tagging

Da für die Korpusanalyse anfänglich noch keine Systeme der automatischen Wortformerkennung mit ausreichender Datenabdeckung existierten, konzentrierte man sich zunächst auf eine rein buchstabenbasierte statistische Analyse. Sie resultiert in Wortformenlisten, welche die Häufigkeiten in bezug auf das Gesamtkorpus und meist auch in bezug auf die einzelnen Genres angeben. Dies illustriert z. B. die Rangliste des Brown-Korpus nach Kučera und Francis, deren Anfang in 4.1 wiedergegeben ist.

Dabei bedeutet z. B. der Eintrag `9543-15-428 HE`, daß die Form HE insgesamt 9 543 mal vorkommt, und zwar in allen 15 Genres, aber nur in 428 der insgesamt 500 Einzeltextproben.

Was in 4.1 aus linguistischer Sicht fehlt, sind grammatische Informationen, insbesondere Bestimmung (i) der Wortart und der Flexionsform (Kategorisierung) und (ii) der Grundform (Lemmatisierung). Um diese Lücke wenigstens teilweise zu schließen, entwickelte N. W. Francis 1980 das System TAGGIT, eine musterbasierte Methode der Kategorisierung, die eine starke manuelle Nachbearbeitung erfor-

[8] Siehe Zip 1932.
[9] Diese Beispiele sind dem Limas-Korpus entnommen.

```
69971-15-500 THE          21341-15-500 IN
36411-15-500 OF           10595-15-500 THAT
28852-15-500 AND          10099-15-485 IS
26149-15-500 TO            9816-15-466 WAS
23237-15-500 A             9543-15-428 HE
```

4.1 Anfang der Frequenzliste im Brown-Korpus

derte. Darauf aufbauend[10] entwickelten Garside, Leech & Sampson 1987 das CLAWS1-System, das versucht, die Kategorisierung aus der Häufigkeitsverteilung der Wortformen zu erschließen. Dieses statistische *tagging* wurde u. a. entwickelt, um schnellere und bessere Ergebnisse bei großen Korpora zu erzielen als mit *pattern matching*.

Statistisches Tagging hat inzwischen große Verbreitung gefunden. Es basiert darauf, daß zunächst die Wortformen eines kleinen Teilkorpus *(core corpus)* in Handarbeit kategorisiert werden – oder eine halbautomatische Kategorisierung zumindest nachträglich sorgfältig ediert und korrigiert wird. Nach dem manuellen Tagging des Teilkorpus werden mit Hilfe von *Hidden Markov Models* (HMMs) die Wahrscheinlichkeiten der Übergänge von einem Tag zum nächsten berechnet. Dann werden die Wahrscheinlichkeiten des manuell getaggten Teilkorpus unter Verwendung eines vereinfachten Tagsets auf das Gesamtkorpus übertragen. Im BNC umfaßt dieses sogenannte *basic (C5) tagset* 61 *labels*.

AJ0	Adjective (general or positive) (e.g. good, old, beautiful)
CRD	Cardinal number (e.g., one, 3, fifty-five, 3609)
NN0	Common noun, neutral for number (e.g. aircraft, data, committee)
NN1	Singular common noun (e.g. pencil, goose, time, revelation)
NN2	Plural common noun (e.g. pencils, geese, times, revelations)
NP0	Proper noun (e. g. London, Michael, Mars, IBM)
UNC	Unclassified items
VVB	The finite base form of lexical verbs (e. g. forget, send, live,return)
VVD	The past tense form of lexical verbs (e. g. forgot, sent, lived, returned)

[10]Siehe Marshall 1987, S. 43 – 5.

[11]Die Verwendungvon HMMs für das grammatische Tagging von Korpora wird z. B. in Leech, Garside & Atwell 1983, Marshall 1983, de Rose 1988, Sharman 1990, Brown, P., V. Della Pietra et al. 1991 beschrieben. Siehe auch K. Church & Mercer 1993.

VVG	The -ing form of lexical verbs (e. g. forgetting, sending, living, returning)
VVI	The infinitive form of lexical verbs (e. g. forget, send, live, return)
VVN	The past participle form of lexical verbs (e. g. forgotten, sent, lived, returned)
VVZ	The -s form of lexical verbs (e. g. forgets, sends, lives, returns)

4.2 Teilmenge des basic *(C5)* tagset

Nachdem das gesamte Korpus auf diese Weise getaggt ist, können – statt reiner Oberflächen – getaggte Wortformen für die Häufigkeitsanalyse verwendet werden. Dabei werden Oberflächen mit verschiedenen Tags als verschiedene Types behandelt. Dies illustriert das folgende Beispiel, das als zufällige Stichprobe aus der getaggten BNC-Liste entnommen wurde, die online zur Verfügung stand.

```
1 activ nn1-np0 1          8 activating aj0-nn1 6
1 activ np0 1              47 activating aj0-vvg 22
2 activa nn1 1             3 activating nn1-vvg 3
3 activa nn1-np0 1         14 activating np0 5
4 activa np0 2             371 activating vvg 49
1 activatd nn1-vvb 1       538 activation nn1 93
21 activate np0 4          3 activation nn1-np0 3
62 activate vvb 42         2 activation-energy aj0 1
219 activate vvi 116       1 activation-inhibition aj0 1
140 activated aj0 48       1 activation-synthesis aj0 1
56 activated aj0-vvd 26    1 activation. nn0 1
52 activated aj0-vvn 34    1 activation/ unc 1
5 activated np0 3          282 activator nn1 30
85 activated vvd 56        6 activator nn1-np0 3
43 activated vvd-vvn 36    1 activator/ unc 1
312 activated vvn 144      1 activator/ unc 1
1 activatedness nn1 1      7 activator/tissue unc 1
88 activates vvz 60        61 activators nn2 18
5 activating aj0 5         1 activators np0 1
```

4.3 Alphabetische Wortformenliste (Stichprobe BNC)

Jeder Eintrag in 4.3[12] besteht erstens aus einer Zahl, welche die Häufigkeit im Gesamtkorpus angibt, zweitens aus der Oberfläche der Wortform, drittens dem Label und viertens der Zahl der Teilkorpora, in denen die Wortform in der angegeben Kategorisierung gefunden wurde. Die verschiedenen Kategorien in 4.3 wurden über ihre Umgebung (Bigramme, Trigramme) im Text statistisch errechnet. Die entsprechende Frequenzliste des BNC besteht aus denselben Einträgen, jedoch nach Häufigkeit statt nach Alphabet geordnet.

4.3 illustriert die Ergebnisse des statistischen Taggers CLAWS4, der für die Analyse des BNC entwickelt wurde und der allgemein als einer der besten statistischen Tagger angesehen wird. Die Fehlerquote[13] von CLAWS4 wird von Leech 1995 auf 1,7 % beziffert, was auf den ersten Blick als sehr gut erscheinen mag.

Man muß jedoch bedenken, daß diese Fehlerquote das Tagging der laufenden Wortformen und nicht der Types betrifft. Angesichts der Tatsache, daß die Abdeckung der letzten 1,2 % der Tokens 83,6 % der Types erfordert (siehe 3.1), kann eine Fehlerrate von 1,7 % auch ein sehr schlechtes Ergebnis bedeuten – nämlich daß über 80 % der Types nicht oder nicht korrekt analysiert werden. Diese Überlegung wird von einer genaueren Betrachtung der Stichprobe 4.3 bestätigt. Für die BNC-Stichprobe 4.3 ergibt sich nämlich eine Fehlerquote von mindestens 60 %.

Zunächst fällt auf, daß von den 38 Einträgen der Stichprobe 27 Einträge mehrfach kategorisiert sind, nämlich `activ` (2), `activa` (3), `activate` (3), `activated` (7), `activating` (6), `activation` (2), `activator` (2) und `activators` (2). Dabei wird der Druckfehler `activ` alternativ als `nn1-np0` und als `np0` klassifiziert, was linguistisch nicht sinnvoll ist. Auch die Klassifikation von `activate` als `np0` ist aus Sicht traditioneller Lexika des Englischen falsch. Der Druckfehler `activatd` wird als `nn1-vvb` kategorisiert, bei `activation.` wird das Interpunktionszeichen nicht eliminiert und der Label `nn0` vergeben, bei `activation/` und `activator/` wird der `/` nicht korrekt interpretiert und der Label `unc` (für *unclassified*) vergeben, wobei die identischen Einträge für `activator/` auch noch separat gezählt werden.

Neben einer hohen Fehlerrate wird die BNC-Statistik durch einen schwachen Präprozessor beeinflußt. Indem etwa verschiedene Zahlen als Wortformen, z. B.

[12]Die getaggten BNC-Listen wurden aus dem WWW genommen (Oktober 1997).

[13]Leider wird weder in Leech 1995 noch in Burnard 1995 spezifiziert, was beim Tagging des BNC als Fehler angesehen wird. Immerhin läuft seit Juni 1995 ein neues Projekt zur Verbesserung des Taggers, ‚The British National Corpus Tag Enhancement Project', dessen Ergebnisse ursprünglich im September 1996 zur Verfügung gestellt werden sollten.

```
1 0.544 crd 1
1 0.543 crd 1
1 0.541 crd 1
```

analysiert werden, resultieren 58 477 zusätzliche Types, was 6,3 % der getaggten BNC-Types entspricht. Weitere Beispiele dieser Art sind Bindestrichsequenzen und Kombinationen von Zahlen mit Maßeinheiten.

Insgesamt wird durch das statistische Labelling die Zahl der Types erheblich aufgebläht. 921 074 getaggten BNC-Types entsprechen z. B. 659 270 Oberflächen-Types. Eine geeignete Behandlung von Zahlen und Bindestrichen würde die Oberflächen-Types um weitere 83 317 auf 575 935 Types reduzieren. Insgesamt wird die Zahl der BNC-Types durch das BNC-Tagging also um mindestens 37,5 % erhöht.

Die Tagging-Analyse des BNC ist ein gutes Beispiel für die Stärken und Schwächen einer *smart solution*. Trotz offensichtlicher Verbesserungsmöglichkeiten bei dem Prä- und Postprozessor unseres konkreten Beispiels verbleiben die folgenden prinzipiellen Grenzen des statistischen Taggings:

1. Die morphosyntaktische Analyse (*Kategorisierung*) der Wortformen ist für die Verwendung durch einen regelbasierten syntaktischen Parser viel zu ungenau.

2. Die Wortformen können nicht auf ihre Grundform zurückgeführt werden (*Lemmatisierung*).

3. Die Wortformen können weder in ihre Allomorphe noch in ihre Morpheme zerlegt werden.

4. Das Gesamtbild der Häufigkeitsverteilungen in einem Korpus wird durch ein künstliches Aufblähen der Typezahl um fast 40 % verzerrt.

4.4 Nachteile des statistischen Taggings

Diese Schwächen treten bei Sprachen mit einer etwas reicheren Morphologie als der des Englischen noch um vieles deutlicher in Erscheinung. Als Vorteile des statistischen Tagging wären dagegen der verhältnismäßig geringe Aufwand und die Robustheit zu nennen, die es nahelegen, statistische Tagger zumindest zur Vorbereitung einer gründlicheren Wortformerkennung zu verwenden.

5 Automatische Wortformerkennung DMM

Die Alternative zum statistischen Tagging ist die *solid solution* einer regel- und lexikonbasierten automatischen Wortformerkennung. Ein solches System ist LA-MORPH, das an der Abteilung für Computerlinguistik der Universität Erlangen-Nürnberg (CLUE) entwickelt wurde. LA-MORPH basiert auf der Grammatiktheorie der linksassoziativen Grammatik[14] und setzt die zu analysierenden Wortformen linksassoziativ (d.h. sukzessive von links nach rechts) aus Allomorphen zusammen. Das Lexikon, in dem diese Allomorphe gespeichert sind, wird vor der Laufzeit automatisch von Allomorphregeln aus einem Grundformlexikon erzeugt.

LA-MORPH arbeitet im Rahmen des an der CLUE entwickelten Programmpakets MALAGA.[15] Größere Anwendungen von LA-MORPH sind die Systeme DMM (Deutsche Malaga-Morphologie)[16], IMM (Italienische Malaga-Morphologie)[17], KMM (Koreanische Malaga-Morphologie)[18] und EMM (Englische Malaga-Morphologie)[19]. Diese Systeme stehen auf der CLUE-Homepage zur Verfügung und können über eine Java™-Schnittstelle[20] getestet werden.

Das Grundformlexikon der DMM besteht im Moment aus circa 49 000 Einträgen in folgender Zusammensetzung:

20 300	Substantive
11 100	Adjektive
10 600	Namen und Akronyme
6 200	Verben
960	Funktionswörter, Partikeln, Suffixe, Präfixe und Präfixoide
49160	Gesamt

Aus diesen knapp 50 000 Grundformen werden regelbasiert circa 65 000 Allomorphe generiert. Aus dem Grundformeintrag für *Haus* werden beispielsweise die Allomorphe *Haus* und *Häus* erzeugt. Insgesamt ergibt sich daraus ein Allomorphiequotient von 1,32 für das Deutsche.

[14] Hausser 1992.
[15] Beutel 1997.
[16] Lorenz 1996.
[17] Wetzel 1996.
[18] Lee 1995.
[19] Leidner 1998.
[20] Knorr 1997.

Beispielsweise würde DMM bei der Wortform Häusermeer zuerst das Allomorph Häus erkennen. Über *lexical lookup* wird die kategorialen Information über das Allomorph Häus und das zugehörige Morphem Haus bestimmt. Als nächstes wird das Allomorph -er- gefunden, analysiert und über eine Regel u. a. als Fuge angehängt.

Die Zusammensetzung Häus/er wiederum läßt nur eine bestimmte Menge möglicher Fortsetzungen (also nachfolgender Regeln und zugehöriger Kategorien nächster Allomorphe) zu. Eine dieser Nachfolgeregeln konkateniert Häus/er mit dem nächsten Allomorph, Meer, das als Substantivstamm des Morphems Meer analysiert wurde. In dieser Weise zerlegt DMM die Wortformen in *Allomorphe* und liefert die entsprechenden *Morpheme*, eine genaue morphosyntaktische Analyse, sowie die *Grundform*.

Falls es mehrere Möglichkeiten gibt, eine Wortform zusammenzusetzen (z. B. Ab/treibung vs. Abt/reibung), so werden diese Analysen morphologisch disambiguiert, d. h. es wird versucht, aufgrund morphologischer Kriterien zu entscheiden, welche der Analysen korrekt ist (oder zumindest korrekter als die anderen). Die Entscheidungskriterien sind hierbei die Art der in der Analyse zusammengesetzten Allomorphe und die Art und Anzahl der Konkatenationsschritte. Außerdem werden morphologisch ambige Analysen, die syntaktisch identisch sind, verschmolzen. Durch morphologische Disambiguierung und Verschmelzung wird erreicht, daß DMM im Schnitt nur 1,05 Analysen pro Wortform liefert (anstatt circa 1,4 Analysen ohne diese Mechanismen).

6 DMM-basierte Analyse des Limas-Korpus

Die automatische Wortformerkennung des DMM-Systems ermöglichte erstmals eine umfassende regelbasierte Analyse des Limas-Korpus. Sie liefert zum einen eine detaillierte morphosyntaktische Charakterisierung der einzelnen Wortformen. Zum anderen ergänzt sie die bekannte Häufigkeitsverteilung der Wort*formen* mit den bisher unbekannten Häufigkeitsverteilungen der *Wörter* (Grundformen), *Morpheme* und *Allomorphe*.

Es folgt zunächst eine Frequenzliste des Limas-Korpus, bei der alle Wortformen, alle Morpheme und alle Allomorphe berücksichtigt werden. Erwartungsgemäß wird der Anfang dieser Frequenzlisten von Funktionswörtern bestimmt, während bei den Morphemen die *gebundenen* Morpheme dominieren.

Rang	Wortformen	Allomorphe	Morpheme
1	39911 der	110423 en	157493 _det_pron
2	39278 die	77607 e	110422 en
3	28898 und	50049 t	77606 e
4	18814 in	39911 der	47317 t
5	12478 den	39278 die	37864 ung
6	11402 von	37860 ung	34479 n
7	11349 das	34479 n	33797 _det
8	11091 zu	31172 er	28898 und_conj
9	9607 des	28949 und	26676 s
10	9466 ist	26676 s	26058 er
11	9175 mit	22103 ein	19335 sein
12	8424 sich	18837 in	18814 in_prepos
13	8108 auf	16125 es	17027 einen
14	8056 nicht	14799 zu	13641 werden
15	7473 im	12478 den	12188 auf_prepos
16	7264 eine	12188 auf	11540 an_prepos
17	7260 sie	11540 an	11402 von_prepos
18	7094 für	11402 von	11091 zu_prepos
19	6716 dem	11349 das	10268 aus_prepos
20	6708 ein	10268 aus	9715 mit_prepos

6.1 Frequenzliste (Limas-Korpus Anfang)

Um die Tabelle nicht durch hochfrequente Funktionswörter (z. B. der) im Bereich der Wortformen und hochfrequente Suffixe im Bereich der Allomorphe und Morpheme (z. B. -en) zu vernebeln und um die problematische Darstellung bestimmter Funktionswörter, z. B. der Artikel der, die, das, dem, den etc., und bestimmter Suffixe, z. B. der Pluralendungen -en, -er, -e, -n etc., als letztlich künstliche Morpheme (z. B. der oder Def-Art bzw. -en oder Plural) zu vermeiden, werden die Frequenzen des Limas-Korpus in 6.2 noch einmal für die *offenen* Wortklassen gezeigt.

Es ist offensichtlich, daß sich Wortformen, Allomorphe und Morpheme in einem Korpus wesentlich besser vergleichen lassen, wenn nur die offenen Wortklassen berücksichtigt werden.

Durch die Beschränkung auf die offenen Wortklassen verringert sich die Zahl der Tokens im Limas-Korpus von 1 059 310 (ohne Interpunktionszeichen, Klam-

Rang	Wortformen	Allomorphe	Morpheme
1	9466 ist	9467 ist	19335 sein
2	5408 werden	5408 werden	12949 werden
3	4125 wird	4126 wird	7513 haben
4	4045 sind	4045 sind	5403 können
5	2817 hat	3552 stell	3531 stellen
6	2505 war	2819 hat	2625 zeit
7	2484 kann	2791 bild	2606 geben
8	1929 haben	2625 zeit	2547 müssen
9	1592 können	2543 kann	2427 bilden
10	1451 wurde	2513 war	2419 nehmen
11	1267 hatte	2309 arbeit	2281 jahr
12	1216 muß	2281 jahr	2272 führen
13	1166 zeit	2273 führ	2183 kommen
14	890 sei	2091 setz	2167 groß
15	890 anderen	2076 ander	2091 setzen
16	835 waren	1978 teil	2079 ander
17	777 soll	1929 haben	2036 gehen
18	767 wurden	1772 neu	1963 sollen
19	755 gibt	1736 bau	1881 lassen
20	719 jahre	1724 komm	1844 sehen

6.2 Frequenzen der offenen Klassen (Limas-Korpus Anfang)

mern etc.) auf 617 952 (den Verhältnissen in 3.1 entsprechend). Die Anzahl der Types verringert sich dagegen nur um 5 186 Types von 98 138 Rängen auf 92 952 Ränge. Von diesen 5 186 Types sind 345 Funktionswörter und 2 100 Zahlen; die restlichen Types sind morphologische Hypothesen, die nicht berücksichtigt werden, weil Information über ihre morphologische Struktur nicht zuverlässig vorliegt.

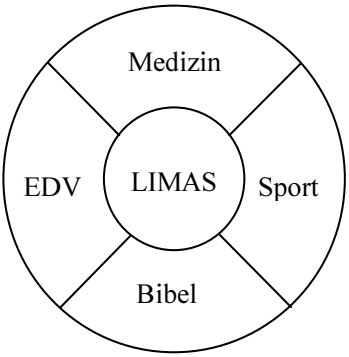
Den 617 952 Tokens bei den Wortformen stehen bei Morphemen und Allomorphen gleichermaßen 713 526 Tokens gegenüber. Die höhere Zahl der Tokens bei Morphemen und Allomorphen resultiert aus der Analyse von Komposita. So wird z. B. Häusermeer als eine Wortform, aber als zwei Morpheme (Haus und Meer) und als zwei Allomorphe (Häus und Meer) gezählt.

Den 92 952 Types der Wortformen der offenen Klassen stehen 21 969 Morpheme und 22 519 Allomorphe gegenüber. Dabei sind von den Wortformen 53 482 oder 57,5 % Hapax Legomena, von Allomorphen und Morphemen sind dagegen nur 7 078 bzw. 6 840 oder jeweils 31,4 % bzw. 31,1 % Hapax Legomena. Die regelbasierte Wortformanalyse reduziert also den Anteil der Hapax Legomena im Vergleich zur traditionellen buchstabenbasierten Methode um beinahe die Hälfte (45,4 %).

Die im Vergleich zu den Wortformen wesentlich kleinere Zahl der Hapax Legomena bei den Morphemen ergibt sich aus der regelbasierten Analyse. Als Wortformen sind z. B. Abbremsung, Babyflaschen und Campingplatz zwar Hapax Legomena im Limas-Korpus, aber ihre Bestandteile bremse, baby, flasche, camping und Platz kommen mehr als einmal vor. Da die Hapax Legomena aus statistischer Sicht als quasi unanalysierbarer Bodensatz eines Korpus betrachtet werden, hat die regelbasierte Wortformanalyse auch den Vorteil, daß sie zu einer wesentlich besseren Ausbeute führt.

7 Domänenspezifische DMM-Analysen

Um einen ersten Eindruck von der Wort- und Morphemverteilung in den niederfrequenten Bereichen des Deutschen zu gewinnen, wurde das Limas-Korpus mit vier domänenspezifischen Korpora gleicher Größe (d. h. je eine Million laufende Wortformen) ergänzt. Diese liegen in den Bereichen Medizin, Sport, EDV und Bibel. Damit ergab sich ein Gesamtkorpus mit folgender Struktur.



7.1 Struktur des CLUE-Korpus

Im CLUE-Korpus dient das Limas-Korpus als relativ hochfrequenter Zentralbereich des deutschen Wortschatzes, da es ja als repräsentatives und balanciertes Korpus konstruiert worden war. Bei den anderen Korpora stellt sich nun die Frage, inwieweit sich ihr Vokabular mit dem des Limas-Korpus und untereinanderüberschneidet.

Die domänenspezifischen Teilkorpora des CLUE-Korpus wurden folgenden Quellen entnommen:

EDV-Korpus:
Als Quelle für das EDV-Korpus dienen Texte aus den Zeitschriften *Computerzeitung*, *iX* und *c't* und den online im Internet verfügbaren Magazinen *InterNet Report* und *Autocad*.
Die Texte der *Computerzeitung* stellen die vollen Jahrgänge 1993 und 1994 dar und lagen auf CD-ROM vor. Ebenfalls auf CD-ROM liegen die Jahrgänge 1994 bis 1996 der Zeitschrift *iX* vor.
Die Texte der Zeitschriften *InterNet Report*[21] und *AUTOCAD-Magazin*[22] des *IWT-Verlags* sowie der Zeitschrift *c't*[23] wurden mit Hilfe des an der CLUE entwickelten Perl-Skriptes `holwww` aus dem WWW (World-Wide Web) beschafft. Diesem Skript wird ein URL (Uniform Resource Locator, „WWW-Adresse") übergeben; das Skript holt dann den Inhalt dieser WWW-Seite und speichert diesen in einer Datei. Die Beschaffung der Daten kann somit weitgehend automatisiert werden.

Sport-Korpus:
Die Quellen für das Sport-Korpus entstammen den Sportteilen der WWW-Versionen der Tageszeitung *Die Welt*. *Die Welt* verfügt über ein Archiv, in dem sich sämtliche Artikel befinden, die in der *Welt Online*[24] seit 17.Mai 1995 erschienen sind. Auf dieses Archiv kann frei zugegriffen werden. Die Artikel der Domäne Sport wurden ebenfalls mit Hilfe des oben genannten Skriptes holwww automatisch beschafft.
Bei der Auswahl der Quellen wurde darauf Wert gelegt, daß die Erscheinungsdaten der Artikel möglichst weit über alle Jahreszeiten verteilt sind, da aufgrund der Natur der Domäne starke saisonale Abweichungen zu erwarten sind.

[21] http://www.iwtnet.de/inet_report/Homepage.html
[22] http://www.iwtnet.de/autocad/Homepage.html
[23] http://www.heise.de/ct/
[24] http://www.welt.de

Medizin-Korpus:
Das Medizin-Korpus entstammt folgenden Quellen: Der Online-Version der *Ärztezeitung*[25], dem WWW-Server des Bundesministeriums für Gesundheit[26], dem Online-Forum des *Instituts für Medizin und Kommunikation*[27], dem Online-Magazin *MedizInfo*[28], dem Online-Magazin *Medizin-Forum*[29] und den WWW-Seiten der Deutschen Herzstiftung[30] und der Deutschen Krebshilfe[31]. Dazu kamen noch eine Reihe von online verfügbaren Fachzeitschriften des Springer-Verlages[32]: *Der Chirurg, Der Hautarzt, Der Internist, Der Nervenarzt, Der Radiologe, Der Schmerz, Der Unfallchirurg* und *Psychotherapeut*. Beschaffung der Texte wie beim EDV-Korpus.

Bibel-Korpus:
Das Bibel-Korpus entstammt zwei Quellen: Der Bibel in der *Elberfelder Übersetzung*[33] und der sog. *Bibel der Häretiker*[34], einer Sammlung von frühchristlichen gnostischen Handschriften. Auch diese Texte wurden automatisch mit holwww beschafft.
Aufgrund begrenzter Ressourcen wurde bei der Zusammenstellung dieser domänenspezifischen Korpora auf eine strenge Synchronizität der Texte sowie eine Randomisierung verzichtet. Dies schien zum einen vertretbar angesichts der in Abschnitt 2 dargestellten Schwierigkeiten bei der Konstruktion wirklich repräsentativer und balancierter Korpora. Zum anderen liegen die Zwecke des CLUE-Korpus (i) im Testen der DMM und (ii) in einer ersten Untersuchung (Machbarkeitsstudie) der Morphemverteilungen in speziellen Domänen.

Beim Testen der DMM am CLUE-Korpus ergaben sich folgende Erkennungsraten:

[25] http://www.aerztezeitung.de/de/htm/net/start/start.htm

[26] http://www.bmgesundheit.de/

[27] http://www.imk.ch/

[28] http://www.medizinfo.com/

[29] http://www.medizin-forum.de/aktuell/

[30] http://www.dsk.de/dhs/aktuell.htm

[31] http://www.krebshilfe.de/

[32] http://www.link.springer.de/link/service/journals/

[33] gopher://wiretap.spies.com/11/Library/Religion/Bible/German

[34] http://www.gwdg.de/~rzellwe/nhs/nhs.html

Korpus	Tokens	erk.	in %	Types	erk.	in %	to/ty
Limas	1236774	1204225	97,37	121650	104106	85,58	10,16
Bibel	1131536	1106629	97,80	37031	29932	80,83	30,56
Sport	1140121	1082154	94,92	64799	50293	77,62	17,59
EDV	1000001	899176	89,92	100208	66975	66,84	9,98
Medizin	1017646	877964	86,28	104425	66421	63,71	9,74
Total	5526079	5170149	93,56	324570	221138	68,14	17,02

7.2 Erkennungsraten der DMM

Bei den laufenden Wortformen (Tokens) liegt die Erkennungsrate der gegenwärtigen DMM zwischen 97,37 % (Limas) und 86,28 % (Medizin). Die Erkennungsrate für das gesamte CLUE-Korpus ist 93,56 %.

Bei den Types der Wortformen liegt die Erkennungsrate zwischen 85,58 % (Limas) und 63,71 % (Medizin). Für das gesamte CLUE-Korpus liegt die Type-Erkennung bei 68,14 %.[35] Bei der Einschätzung dieser Werte sollte die in 3.1 dargestellte Korrelation von Types und Tokens in Korpora im Auge behalten werden.

Es zeigt sich, daß bei den domänenspezifischen Korpora die Type-Erkennungsrate mit dem Token/Type-Verhältnis (to/ty) korreliert. Je weniger Tokens es zu einem Type gibt, je weniger oft also eine Wortform wiederholt wird, desto höher ist die Anzahl der Types im Korpus – was sich entsprechend auf die Type-Erkennungsrate auswirkt.

Eine Untersuchung der Vokabularüberschneidung[36] verschiedener Teilkorpora ergibt eine Fülle möglicher Morphemklassen, die n-Listen genannt werden, wobei n für die Namen der Teilkorpora steht. Beim CLUE-Korpus gibt es z. B. folgende n-Listen:

Morpheme, die in allen 5 Teilkorpora vorkommen
Morpheme, die jeweils in nur 4 Teilkorpora vorkommen
Morpheme, die jeweils in nur 3 Teilkorpora vorkommen
Morpheme, die jeweils in nur 2 Teilkorpora vorkommen
Morpheme, die jeweils in nur 1 Teilkorpus vorkommen

[35]Daß dieser Wert nicht dem Mittel der einzelnen Erkennungsprozente entspricht, liegt an der Vokabularüberschneidung zwischen den Teilkorpora.

[36]Siehe Schwarz 1996.

In jeder dieser n-Listen wird die Frequenz der Morpheme relativ zu den betrachteten Teilkorpora angegeben. Dies zeigt die folgende Tabelle aus Morphemen, die ausschließlich in den Domänen EDV und Medizin vorkommen, geordnet nach ihrer Gesamthäufigkeit in beiden Domäne, für die Ränge 1 – 20.

Rang	Morphem	Gesamt	EDV	Medizin
1	datei	951	950	1
2	radiologisch	184	1	183
3	interaktiv	172	167	5
4	frame	170	169	1
5	editor	165	161	4
6	spezifikation	160	159	1
7	insulin	154	19	135
8	diabetes	149	10	139
9	joint	142	8	134
10	prospektiv	128	1	127
11	infusion	114	1	113
12	macintosh	103	100	3
13	disk	99	97	2
14	neuronal	98	39	59
15	expression	94	1	93
16	environment	93	90	3
17	skript	92	90	2
18	pixel	84	73	11
19	zertifizieren	82	81	1
20	array	81	80	1

7.3 Morpheme und normierte Frequenzen zweier Domänen

Beispielsweise kommt das Morphem Datei in den Teilkorpora Limas, Sport und Bibel nicht vor, wohl aber in den Teilkorpora EDV und Medizin. Aufgrund seiner unterschiedlichen Häufigkeit (950 vs. 1) ist es für die Domäne EDV wesentlich charakteristischer als für Medizin. Entsprechend ist es mit dem Morphem radiologisch, das aufgrund seiner Häufigkeit charakteristischer für EDV ist als für Medizin.

Zum Schluß ein Vergleich der Morpheme, die in jeweils nur einem der vier domänenspezifischen Teilkorpora vorkommen.

	EDV		Medizin		Bibel		Sport	
1	prozessor	541	lymphom	547	jakobus	167	steffi	512
2	raid	206	maligne	412	ephraim	151	klinsmann	429
3	modem	156	endothel	186	sündigen	144	wimbledon	325
4	proprietär	89	paracetamol	144	zion	144	sammer	255
5	modular	45	laparoskop-	130	joab	133	villeneuve	203
6	debi	40	suppressiv	119	moab	120	berti	202
7	borchers	31	median	113	jonatan	114	sampras	184
8	explorer	29	palliativ	113	samaria	113	hoeneß	182
9	megabit	29	inzidenz	98	knechten	110	hässler	179
10	paperback	28	hypertonie	93	gilead	106	hunke	131
11	portiere	27	ruptur	93	pleroma	105	kirsten	125
12	ergonomisch	22	seehofer	91	absalom	102	derby	113
13	postleitzahl	22	poliklinik	88	esau	98	köpke	100
14	multiplex	20	mortalität	86	jerobeam	97	agassi	98
15	integer	18	psychosozial	81	pharisäer	97	doping	97
16	platine	18	zyste	81	ahab	91	ottmar	89
17	drda	17	septisch	80	elia	91	strunz	81
18	assembler	16	radiologe	76	edom	86	babbel	78
19	permission	16	fixateur	75	nebukadnezar	86	edberg	76
20	synergie	16	zerebral	73	joschafat	84	hertha	71

7.4 Domänenspezifische Unique-Vokabulare (Morpheme)

Es zeigt sich, daß die Teilkorpora mit einem hohen Token/Type-Verhältnis (Bibel und Sport) in ihrem Unique-Vokabular einen hohen Anteil an Eigennamen haben, die zudem häufig vorkommen. Die n-Listen des CLUE-Korpus stehen in ihrer Gesamtheit auf dem CLUE-Web-Server zur Verfügung.

8 Conclusio

Die von der DMM gelieferten Morpheme sind gewissermaßen ein Nebenprodukt einer regelbasierten Wortformerkennung, deren eigentliches Ziel eine präzise morphosyntaktische Analyse für das syntaktische Parsen ist. Im Bereich der Korpusanalyse zeigt sich jedoch, daß die Häufigkeitsverteilungen auf der Ebene der Morpheme ein wesentlich klareres Bild von einer Sprache geben als auf der

Ebene der Wortformen.

Neben einer theoretischen Untersuchung des deutschen Wortschatzes in verschiedenen Domänen haben die hier beschriebenen Methoden auch praktische Anwendungen. Zum einen konnte gezeigt werden, daß bei einer flektierenden Sprache wie dem Deutschen mit einer DMM-basierten Suche eine Recall/Precision-Verbesserung zwischen 42,9 % und 110,5 % erreicht werden konnten.[37] Zum anderen liegt es nahe, die domänenspezifischen n-Listen zur automatischen Klassifikation von Texten zu verwenden.

Bibliographie

Bergenholtz, H. (1976) *„Zur Morphologie deutscher Substantive, Verben und Adjektive. Probleme der Morphe, Morpheme und ihrer Beziehungen zu den Wortarten.“* In: Alfred Hoppe (Hrsg): Beihefte zur kommunikativen Grammatik. Bonn.

Bergenholtz, H. (1989) *“Korpusproblematik in der Computerlinguistik. Konstruktionsprinzipien und Repräsentativität.“* In: Hugo Steger (Hrsg.): Handbücher zur Sprach- und Kommunikationswissenschaft (Bd. IV). Berlin, New York 1989.

Beutel, B. (1997) *Malaga 4.0*, CLUE-Manual.

Brown, P., S. Della Pietra, V. Della Pietra and R. Mercer (1991) *“Word sense disambiguation using statistical methods”,* in: Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, Berkeley, CA, June 1991, 264-270.

Burnard, L. (ed.) (1995) *Users Reference Guide British National Corpus Version 1.0*, Oxford University Computing Services.

Church, K. & R. L. Mercer (1983) *“Introduction to the special issue on computational linguistics using large corpora.”* Computational Linguistics, Vol. 19:1, 1 – 24.

DeRose, S. (1988) *“Grammatical category disambiguation by statistical optimization.”* Computational Linguistics, 14:1, 31 – 39.

Garside, R., G. Leech und G. Sampson (1987) *The Computational Analysis of English*, Longman, London & New York.

Francis, W. N. (1980) *“A tagged corpus: Problems and prospects,”* in: S. Greenbaum, G. Leech and J. Svartvik (eds.) 1980, pp. 192 – 209.

Francis, W. N. und H. Kučera (1982) *Frequency Analysis of English Usage: Lexicon and Grammar*, Houghton Mifflin, Boston.

[37] Piotrowski 1998

Hausser, R. (1992) *"Complexity in Left-Associative Grammar"*, Theoretical Computer Science, Vol. 103, Elsevier.

Hausser, R. (ed.) (1996) *Linguistische Verifikation*. Dokumentation zur Ersten Morpholympics, Max Niemeyer Verlag, Tübingen.

Hess, K., J. Brustkern und W. Lenders (1983) *Maschinenlesbare deutsche Wörterbücher*, Max Niemeyer Verlag, Tübingen.

Hofland, K. und S. Johansson (1980) *Word Frequencies in British and American English*, London: Longman.

Knorr, O. (1997) *Entwicklung einer JAVA-Schnittstelle für Malaga*, CLUE-betreute Studienarbeit der Informatik.

Kučera, H. & W. N. Francis (1967) *Computational Analysis of Present-day English*, Brown University Press, Providence, Rhode Island.

Garside, R., G. Leech & G. Sampson (1987) *The Computational Analysis of English*, Longman, London & New York.

Leidner, J. (1998) *Linksassoziative morphologische Analyse des Englischen mit stochastischer Disambiguierung*, CLUE-Magisterarbeit.

Lee, K. (1995) "*Recursion Problems in Concatenation: A Case of Korean Morphology"*, Proceedings of PACLIC 10, the 10th Pacific-Asian Conference on Language, Information and Computation.

Leech, G. (1995) *"A brief user's guide to the grammatical tagging of the British National Corpus"*, Web-Seite.

Leech, G., R. Garside & E. Atwell (1983) *"The automatic grammatical tagging of the LOB Corpus"*, ICAME Journal 7, 13 – 33.

Lenders, W. und G. Willee (1986) *Linguistische Datenverarbeitung*, Westdeutscher Verlag, Opladen.

Lorenz, O. (1996) *Automatische Wortformerkennung für das Deutsche im Rahmen von Malaga*, CLUE-Magisterarbeit.

Marshall, I. (1983) *"Choice of grammatical word-class without global syntactic analysis: Tagging words in the LOB Corpus"*, Computers and the Humanities, Vol. 17, 139 – 150.

Marshall, I. (1987) *"Tag selection using probabilistic methods"*, in Garside et al. (eds.).

Meier, H. (1964) *Deutsche Sprachstatistik*. Erster Band. Hildesheim.

Oostdijk, N. (1988) *"A corpus linguistic approach to linguistic variation"*, in G. Dixon (ed.): Literary and Lingustic Computing, Vol. 3.1.

Oostdijk, N. & P. de Haan (1994) *Corpus-based Research into Language*. Editions Rodopi B. V., Amsterdam-Atlanta, GA.

Piotrowski, M. (1998) *NLP-Supported Full-text Retrieval*, CLUE-Magisterarbeit.

Sharman, R. (1990) *Hidden Markov Model Methods for Word Tagging*, Report 214. Winchester: IBM UK Scientific Centre.

Schwarz, R. (1996) *Dynamische Aktivierung domänenspezifischer Teillexika*, CLUE-Magisterarbeit.

Wetzel, C. (1996) *Erstellung einer Morphologie für Italienisch in Malaga*, CLUE-betreute Studienarbeit der Informatik.

Zierl, M. (1997) *Ein System zur effizienten Korpusspeicherung und -abfrage*, CLUE-Magisterarbeit.

Zipf, G.K. (1932) *Selected Studies of the Principle of Relative Frequency in Language*, Oxford.

Kryptographiebasierte Kommunikationsformen für Vereine und Verbände

Bettina Mielke
Juristische Fakultät
Universität Regensburg
bettina.mielke@jura.uni-regensburg.de

Christian Wolff
Institut für Informatik
Universität Leipzig
wolff@informatik.uni-leipzig.de

1 Einführung

Die Verfügbarkeit unterschiedlicher Kommunikationsformen innerhalb des Internet wirkt sich bereits seit einiger Zeit auch auf die Arbeit von Vereinen und Verbänden aus:

- Informationen über Verbandsaktivitäten werden über das WWW bereitgestellt.
- Mitglieder werden über elektronische Postverteiler per e-mail informiert oder
- können in Diskussionsforen über unterschiedliche Themengebiete mit Bezug zur Arbeit des Vereins oder Verbands durchführen.

In diesem Aufsatz wollen wir aufzeigen, wie unter Einsatz kryptographischer Verfahren weitere typische Kommunikationsformen (Meinungsbildungsprozesse, Umfragen, Wahlen) mit elektronischen Mitteln durchgeführt werden können und welche rechtlichen Rahmenbedingungen dabei zu beachten sind. Die Untersuchung geht dabei von der Situation der *Gesellschaft für linguistische Datenverarbeitung* (GLDV) als typischem Beispiel einer wissenschaftlichen Vereinigung aus.

1.1 Rahmenbedingungen elektronischer Kommunikationsformen

Der Ansatzpunkt für den Einsatz elektronischer Kommunikationsformen über das bereits oben angedeutete Maß hinaus ergibt sich für überregional (oder international) tätige Vereine und Verbände aus folgenden Merkmalen:

- Ein räumlich verstreuter heterogener Mitgliederkreis.
- Die Mitglieder haben zumeist Zugang zum Internet und seinen Diensten (e-mail, WWW) und die Realisierung bzw. Nutzung elektronischer

Kommunikationsformen ist mit geringem Kostenaufwand möglich, bei wissenschaftlichen Vereinigungen kann i. d. R. die technische Infrastruktur der Hochschulen genutzt werden.

- Traditionelle Kommunikationsverfahren bei Wahlen oder Abstimmungen sind zeit- und kostenintensiv.
- Gerade größere Verbände können über Mitgliederversammlungen nur einen Bruchteil der Mitglieder erreichten – im Fall der GLDV kann man annehmen, daß bei den Mitgliederversammlungen weniger als 20 % der Mitglieder anwesend sind. Es ist wünschenswert, aber mit traditionellen Mitteln zu aufwendig, *alle* Mitglieder eines Verbands in Meinungsbildungs- und Beschlußfassungsprozesse einzubinden.
- Mitgliederversammlungen finden – nicht zuletzt auch wegen des damit verbundenen Aufwands – in zeitlich relativ großen Abständen statt (im Fall der GLDV je nach Anbindung an unterschiedliche Fachtagungen bis zu 19 Monate). Zumindest für die satzungsgemäß der Mitgliederversammlung vorbehaltenen Entscheidungen zieht das eine nicht unbeachtliche Trägheit der Entscheidungsprozesse nach sich.

Zwar lassen sich aus diesen Beobachtungen Argumente für eine weitgehende Verwendung elektronischer Kommunikationsformen für zentrale Verbandsaufgaben wie Wahlen oder Abstimmungen ableiten; zu beachten ist jedoch, daß auch mittelfristig nicht davon ausgegangen werden kann, daß alle Mitglieder etwa an elektronischen Wahlverfahren teilhaben können oder wollen. Alle Realisierungsvorschläge müssen daher *ergänzenden Charakter* aufweisen, der erst langfristig auf eine Ablösung der traditionellen Kommunikationsmittel abzielt.

1.2 Sicherheitsrelevante Voraussetzungen elektronischer Kommunikation

Will man über rein informelle oder informative Dienste im Internet wie e-mail oder das WWW hinausgehen, stellt sich die Frage, wie wesentliche Sicherheitskriterien gewährleistet werden können. Zu ihnen gehören u. a.

- die Authentisierung (*authentication*) der Kommunikationspartner,
- die Vertraulichkeit (*confidentiality*) der Kommunikation, insbesondere die Gewährleistung der Anonymität von Wahlen und
- die Integrität (*integrity*) der kommunizierten Information.[1]

Neben diese zentralen Kriterien sicherer Kommunikation treten zusätzliche Faktoren, die bei der Operationalisierung kryptographischer Verfahren eine Rolle spielen, so etwa die Transparenz des Verfahrens und die Vertrauenswürdigkeit einer Anwendung.

1.3 Technische Umsetzung

Um die oben genannten Sicherheitskriterien erfüllen und eine rechtlich zulässige Alternative zu den traditionellen Kommunikationsformen anbieten zu können, bedarf es kryptographischer Verfahren, mit denen

- sich die Identität eines Kommunikationspartners feststellen läßt,
- Anonymität etwa bei geheimen Wahlen gewährleistet ist,
- Information so verschlüsselt werden kann, daß sie unbefugte Dritte nicht entziffern können.

Bei kryptographischen Verfahren wird die zu übermittelnde Nachricht (der *Klartext, plain text*) mit Hilfe einer Umwandlungsvorschrift (des Kryptoalgorithmus) in *Geheimtext* (*cyphertext*) umgewandelt. Neben *symmetrischen* Verfahren, bei denen Sender wie Empfänger über denselben (geheimen) Schlüssel zur Verschlüsselung bzw. Entzifferung verfügen, haben sich in den letzten Jahren vor allen sog. *asymmetrische* Verfahren durchgesetzt, bei denen der Klartext mit Hilfe eines *öffentlichen* Schlüssels verschlüsselt und mit Hilfe eines *privaten* Schlüssels, der nur seinem Inhaber bekannt und durch eine Paßwortphrase o. ä. geschützt ist, entziffert werden kann.[2]

Auf der Basis asymmetrischer Verfahren lassen sich auch *digitale Signaturen*, das elektronische Pendant zur eigenhändigen Unterschrift realisieren: Der Unterzeichner signiert ein Dokument mit Hilfe seines privaten Schlüssels. Unter Zuhilfenahme des öffentlichen Schlüssels kann der Empfänger verifizieren, daß ein Dokument von einem bestimmten Sender stammt und daß es *wie unterzeichnet* bei ihm eingetroffen ist, d. h. daß keine Modifikationen am Dokument vorgenommen worden sind. Dabei ist vorausgesetzt, daß der öffentliche Schlüssel in geeigneter Form publiziert worden ist (auf einem frei zugänglichen *key server* oder z. B. auch als ASCII-Text auf der Homepage des Schlüsselinhabers). Die Umsetzung solcher kryptographischer Techniken kann grundsätzlich über Standardsoftware (z. B. das für nicht-kommerzielle Zwecke frei erhältliche *Pretty Good Privacy – PGP*[3]) oder über proprietäre Eigenentwicklungen unter Verwendung kryptographischer *application programming interfaces* (APIs) erfolgen (s. u. Kap. 3).

1.4 Ansatzpunkte für kryptographische Verfahren

Für die Einführung kryptographischer Techniken lassen sich ausgehend von den bewährten traditionellen und elektronischen Kommunikationsformen im Vereins- und Verbandsleben im wesentlichen folgende Ansatzpunkte erkennen:

1) Durchführung von (Vorstands- und Beirats-) Wahlen mit elektronischen Mitteln.
2) Meinungsbildung und Beschlußfassung durch elektronische Kommunikation in Ergänzung der Rolle einer Mitgliederversammlung.
3) Einführung neuer Vereinsorgane durch Satzungsänderung, deren Funktionsweise ganz wesentlich vom Einsatz elektronischer Kommunikationsformen geprägt ist.

Im Mittelpunkt dieses Aufsatzes steht die rechtliche Bewertung elektronischer Vorstandswahlen vor dem Hintergrund des deutschen Vereinsrechts (Kap. 2) sowie ihre praktische Umsetzung und Implementierung (Kap. 3). Zu den Vorschlägen 2) und 3) erfolgen in Kap. 4 einige Hinweise.

2 Rechtliche Aspekte

2.1 Vereinsrechtliche Voraussetzungen

Die Verfassung eines rechtsfähigen Vereins wird gemäß § 25 BGB durch die gesetzlichen Vorschriften sowie durch die Satzung bestimmt. Hinsichtlich der gesetzlichen Vorschriften ist zu unterscheiden, ob es sich um zwingende oder um dispositive, d.h. durch die Satzung abänderbare, Vorschriften handelt. Aus § 40 BGB ergibt sich, welche der Vorschriften dispositiv sind und im Umkehrschluß daraus, welche Bestimmungen zwingender Bestandteil der Vereinsverfassung sind. Soweit die Satzung keine vom Gesetz abweichende Regelung trifft, sind die §§ 26 – 39 BGB und die nicht abänderbaren Vorschriften über die Auflösung des Vereins (§ 41 BGB) und über die Liquidation des Vereinsvermögens (§§ 47 – 52 BGB) Bestandteile der Verfassung jeden Vereins[4].

Die Verfassung eines Vereins wird weiter durch die Regelungen der Vereinssatzung festgelegt. Ihren Inhalt können die Gründer und später die Mitgliederversammlung durch Satzungsänderungsbeschluß im Rahmen der Privatautonomie frei gestalten.[5] Beim eingetragenen Verein stellen die §§ 57, 58 BGB die Mindesterfordernisse für den Satzungsinhalt auf. Hinsichtlich der hier zu besprechen-

den Vorschläge geben § 58 Nr. 3 und 4 BGB an, daß die Satzung Bestimmungen über die Bildung des Vorstands und die Voraussetzungen, unter denen die Mitgliederversammlung zu berufen ist, sowie über die Form der Berufung und die Beurkundung der Beschlüsse zu enthalten hat.

Nach der Rechtsprechung des Bundesgerichtshofs müssen darüber hinaus „die das Vereinsleben bestimmenden Grundentscheidungen" in der Satzung enthalten sein.[6] Da eine Satzungsänderung nur durch Mehrheitsbeschluß und Eintragung in das Vereinsregister vorgenommen werden kann, genießt der Inhalt der Satzung einen gewissen „Bestandsschutz"[7], der zum Schutz der Minderheit und der einzelnen Mitglieder führt.[8] Um diese Schutzfunktion zu gewährleisten, müssen die satzungsmäßigen Regelungen eine hinreichende Regelungsdichte aufweisen.[9] Während die Grundsatzentscheidungen in der Satzung festzulegen sind, können Einzelheiten einer näheren Regelung vorbehalten sein; sie können in nachrangigen Ordnungen, etwa Verfahrensordnungen für Vereinsgerichte, Benutzungsordnungen für Vereinsanlagen[10] oder wie in unserem Fall der Wahlordnung ausgestaltet sein. Solche nachrangigen Ordnungen müssen von dem in der Vereinsverfassung für zuständig erklärten Organ aufgrund einer Ermächtigung erlassen sein und dürfen nicht gegen die Satzung verstoßen.[11]

Insofern ist also bei allen die Organisation eines Vereins betreffenden Regelungen zu fragen, ob eine entsprechende Bestimmung gegen zwingendes Recht verstößt, ob sie einer Bestimmung in der Satzung widerspricht und schließlich ob sie als eine das Vereinsleben bestimmende Grundsatzentscheidung in die Satzung aufzunehmen ist oder auch in einer nachrangigen Neben- oder Vereinsordnung geregelt werden darf.

2.2 Stellung digitaler Signaturen nach dem Signaturgesetz

Das Signaturgesetz (SigG, = Artikel 3 des Informations- und Kommunikationsdienstegesetz, IuKDG) regelt den Einsatz digitaler Signaturen als gesetzlich geregeltem Sicherheitsstandard. Es sind mit der Einführung digitaler Signaturen aber keine Rechtsfolgen verbunden.[12] GEIS bezeichnet das SigG als „gesetzgeberischen Torso"[13], da es die Signatur eben *nicht* dem Schriftformerfordernis des BGB oder der Privaturkunde des § 416 ZPO gleichstellt. Im konkreten Fall bedeutet dies, daß nicht davon auszugehen ist, daß elektronische Kommunikationsformen unter Einsatz kryptographischer Algorithmen ohne weiteres *gesetzlich vorgeschriebenen* schriftlichen Verfahren äquivalent sind oder diese verdrängen können. Im Vereinsrecht spielt dieser Gesichtspunkt jedoch nur eine untergeordnete Rolle, da fast alle Bereiche durch die Satzung frei ausgestaltet werden können (s. o.).[14]

2.3 Durchführung von Vorstandswahlen

Wie die Wahl des Vorstands zu erfolgen hat, ist gesetzlich nicht zwingend festgelegt. § 27 Abs. 1 BGB bestimmt, daß die Bestellung des Vorstandes durch Beschluß der Mitgliederversammlung erfolgt, falls die Satzung nichts anderes vorschreibt (§ 40 BGB).[15] Die Satzung der GLDV trifft hier jedoch eine andere Bestimmung: Gemäß § 19 der Satzung der GLDV wird die „Wahl des Vorstands und des Beirats als Briefwahl [...] durchgeführt". Zum genauen Ablauf der Briefwahl äußert sich die Satzung nicht. In Satz 4 von § 19 der Satzung heißt es nur: „Alles Nähere regelt die Wahlordnung". In der Wahlordnung ist zum technischen Ablauf lediglich ausgeführt, daß „in Briefwahl geheim gewählt" wird. Eine nähere Bestimmung des Begriffs *Briefwahl* erfolgt nicht – weder in der Satzung, noch in der Wahlordnung.

Da die Satzung auf Dauer angelegte Regeln für eine Vereinigung mit wechselndem Mitgliederbestand schafft, ist sie nach zutreffender Ansicht aus sich heraus auszulegen, d. h. die Auslegung kann sich nicht an dem Willen oder den Interessen der Gründer orientieren.[16] In erster Linie ist somit neben Sinn und Zweck der Regelung[17] der Wortlaut maßgebend, und zwar „in einem durch den allgemeinen Sprachgebrauch, evtl. auch durch die Fachsprache in bestimmten Lebensbereichen festgelegten Sinn"[18]. Bei der Auslegung des Begriffs *Briefwahl* wird man sich an anderen Bestimmungen, die eine Briefwahl vorsehen, orientieren können. So regeln z. B. § 36 Bundeswahlgesetz i. V. m. §§ 66, 74 f. Bundeswahlordnung die Briefwahl. Wesentliches Element dieser Bestimmungen ist, daß der Wähler dem Wahlleiter in einem verschlossenen Wahlbriefumschlag seinen Wahlschein und in einem besonderen verschlossenen Umschlag seinen Stimmzettel übersendet. Damit ist gewährleistet, daß der abgegebene Stimmzettel von einem berechtigten Wähler stammt und die Stimmabgabe selbst anonym erfolgt. Diese wesentlichen Bestandteile, wie im nachfolgenden Abschnitt beschrieben, lassen sich durch elektronische Verfahren sicherstellen. Eine Wahl in Form einer „traditionellen" Briefwahl ist also nicht notwendig, um diesen beiden Anforderungen gerecht zu werden. Da die Satzung den Ablauf der Briefwahl der Wahlordnung überläßt, könnte eine Änderung der Wahlordnung ausreichen. In sie wäre etwa einzufügen, daß die Briefwahl *auch* elektronisch durchgeführt werden kann. Selbstverständlich müssen dabei Mitglieder, die keinen Internetzugang haben, die Möglichkeit haben, sich weiterhin durch „traditionelle" Briefwahl zu beteiligen. Dies gebietet bereits der ungeschriebene Rechtsgrundsatz der Gleichbehandlung der Mitglieder.[19] Ob zusätzlich eine Änderung der Satzung notwendig ist, hängt im wesentlichen davon ab, ob man diese Bestimmung als für das Vereinsleben so

wesentliche Grundentscheidung ansieht, daß sie einer satzungsmäßigen Festlegung bedarf.

Zu den Grundentscheidungen zählen die Regelungen von Zweck und Mitteln des Vereins, von Voraussetzungen und Folgen der Mitgliedschaft, von Bildung, Bestellung und Wirkungskreis der Organe sowie von Sitz und Namen.[20] Die Abgrenzung zwischen den in die Satzung aufzunehmenden Grundentscheidungen und den außerhalb der Satzung in Nebenordnungen regelbaren Vereinsangelegenheiten ist im Einzelfall schwierig und umstritten.[21] Wenn man den Sinn der Satzungsvorschrift hinsichtlich der Briefwahl darin sieht, abweichend von der dispositiven Regelung des § 27 BGB und der in der Vereinspraxis gängigsten Art der Vorstandswahl[22] die Wahl des Vorstands ohne persönliche Anwesenheit auf der Mitgliederversammlung durchzuführen, muß man davon ausgehen, daß die nähere Ausgestaltung des technischen Ablaufs dieser Wahlmöglichkeit – ebenso wie bereits jetzt – der Wahlordnung überlassen werden kann. Dies gilt insbesondere dann, wenn die entscheidenden Anforderungen, die an eine Briefwahl zu stellen sind (vgl. oben) durch elektronische Mittel gewährleistet werden und die elektronische Wahl nur eine zusätzliche Möglichkeit der Stimmabgabe – neben der „traditionellen“ Briefwahl – darstellt.

3 Elektronische Vorstandswahl

Das nachfolgend beschriebene Verfahren stellt die elektronische Umsetzung einer Vorstandswahl auf der Basis asymmetrischer Kryptographie dar. Es wird anschließend an voranstehende rechtliche Würdigung angenommen, daß

1) die elektronische Form durch Änderung der Wahlordnung (Beschluß der Mitgliederversammlung) sanktioniert ist,
2) die an der elektronischen Wahl teilnehmenden Mitglieder über e-mail und WWW-Anschluß verfügen und
3) für alle anderen Mitglieder weiterhin die schriftliche Wahlform zur Verfügung steht.

Für jedes Vereinsmitglied wird ein Schlüsselpaar generiert; der Verein fungiert dabei als eine Art Zertifizierungsstelle[23], wobei die Schlüssel ausschließlich für die verbandsinterne Kommunikation geeignet sind. Jedes Mitglied erhält den öffentlichen Schlüssel seines Schlüsselpaars per e-mail zugesandt. Mit dem Schlüssel kann der Wähler den Wahlschein verschlüsseln. Der Verein verwaltet die

privaten Schlüssel, um Nachrichten (hier: den Wahlschein) verifizieren zu können (*Authentisierung*). Zusätzlich generiert die Wahlsoftware mit Hilfe eines symmetrischen Verfahrens einen Schlüssel, mit dem der Wahlzettel *vor* der Verschlüsselung durch den Mitgliedsschlüssel verschlüsselt wird. Es kommt also in Analogie zur Briefwahl ein zweistufiges Verfahren zum Einsatz:

1) Das Mitglied füllt den Wahlzettel aus, die Wahlsoftware verschlüsselt ihn mit Hilfe des symmetrischen Schlüssels („Verbandsschlüssel"), ohne daß der Benutzer eingreifen müßte.
2) Der verschlüsselte Wahlzettel wird nochmals mit dem mitgliedsbezogenen öffentlichen Schlüssel verschlüsselt.

Die so entstandene Nachricht wird an den Wahlleiter gesandt. Dieser kann anhand der verfügbaren öffentlichen Schlüssel zunächst feststellen, daß der Wahlzettel von einem stimmberechtigten Mitglied stammt und den verbleibenden, mit dem Verbandsschlüssel verschlüsselten Wahlzettel in die (elektronische) Wahlurne geben. Damit ist die Anonymität der Wahl bei gleichzeitiger Überprüfung der Wahlberechtigung gewährleistet. Nach Ablauf der Wahlfrist können alle eingegangenen anonymisierten und verschlüsselten Wahlzettel entschlüsselt und ausgezählt werden.

3.1 Implementierung

Wie bereits angedeutet, bieten sich im wesentlichen zwei Szenarien für die Implementierung des angedeuteten Verfahrens:

Szenario I - Einsatz von Standardsoftware
Setzt man ein geeignetes Kryptographiepaket (z. B. *Pretty Good Privacy*) bei allen Teilnehmern voraus, so könnte das Verfahren wie folgt ablaufen:

1. Generieren der Schlüsselpaare mit *PGP* durch den Verband
2. Verteilen der öffentlichen Schlüssel per e-mail an die Mitglieder
3. Verteilen des Wahlzettels als e-mail-Formular
4. Ausfüllen und (zweimaliges) Verschlüsseln des Wahlzettels mit *PGP*
5. Rücksenden per e-mail
6. Verifikation der Wahlberechtigung (1. Entschlüsselung)
7. Entschlüsseln der Wahlzettel durch den Wahlleiter (2. Entschlüsselung)
8. Auswertung

Dieses Szenario hat den wohl entscheidenden Nachteil, daß die Installation eines komplexen Softwarepakets nicht nur einen organisatorischen Zusatzaufwand darstellt, sondern auch eine zusätzliche Hemmschwelle darstellt, da sich jeder Einzelne in Bedienung und Funktionsweise einarbeiten muß. Demgegenüber versucht das nachfolgende Szenario – um den Preis eines höheren Entwicklungsaufwands – eine Lösung aufzuzeigen, die Benutzungsfreundlichkeit mit sehr wenigen notwendigen Interaktionsschritten zu verbinden sucht:

Szenario II – Webbasierte Eigenentwicklung
Bei dieser Lösung steht dem Nutzer für die Wahl ein Wahlzettel als Java™-Applet zur Verfügung, den er in einem Browser von der Verbands-WebSite laden, ausfüllen und versenden kann:

1. Generieren der Schlüssel mit Hilfe des *Java™ Cryptography-API*
2. Verteilen der öffentlichen Schlüssel per e-mail an die Wähler
3. Ausfüllen, (zweimaliges) Verschlüsseln und Signieren des Wahlzettels
4. Abschicken an den Vereins-Webserver
5. Verifikation der Wahlberechtigung (1. Entschlüsselung)
6. Entschlüsseln der Wahlzettel durch den Wahlleiter (2. Entschlüsselung)
7. Auswertung

Der Weg einer Eigenentwicklung hat den zusätzlichen Vorteil, daß sich auch die Schritte 5 – 7 weitgehend automatisieren lassen. Deshalb haben wir uns für dieses Szenario entschieden. Die Programmiersprache Java™ wird verwendet, da sie über ein geeignetes kryptographisches API verfügt, die *Java Cryptography Architecture* (JCA)[24], und gleichzeitig die Realisierung von *active contents* im WWW erlaubt. Die *JCA* gliedert sich in zwei Teile:

- Einen allgemein verfügbaren Teil, der ein plattform- und algorithmenneutrales Programmierinterface für die Entwicklung kryptographischer Anwendungen sowie Basisklassen für die Erstellung digitaler Signaturen und sog. *Message Digests* enthält und
- die sog. *Java Cryptography Extension* (JCE), die darüber hinaus Algorithmen für die Verschlüsselung von Daten bereitstellt.

Aufgrund der U.S.-amerikanischen Exportrestriktionen dürfen die Java™ *JCEs* von SUN nicht exportiert werden, da nach geltendem amerikanischen Recht (*Ex-*

port Control Act, Arms Export Control ACT etc.) Datenverschlüsselungsverfahren als *Waffen* (sic!) gelten.[25] Deshalb wird auf die Reimplementierung der Java™ *JCEs* der TU Graz zurückgegriffen.[26]

3.2 Die eingesetzten kryptographischen Algorithmen

Der Klartext, d.h. der eigentliche Stimmzettel wird mit einem symmetrischen Verfahren verschlüsselt. Hierbei kommt der *International Data Encryption Standard* (*IDEA*) zum Einsatz, das nach allgemeiner Ansicht derzeit mächtigste kryptographische Verfahren.[27] *IDEA* arbeitet als sog. Blockchiffrierverfahren auf der Basis der Mischung von Operationen unterschiedlicher algebraischer Gruppen, d. h. über je 64 Bit Klartext werden die Operationen XOR, Addition Modulo 2^{16} und Multiplikation Modulo $2^{16} + 1$ (eine Primzahl) eingesetzt. Der verschlüsselte Text sowie der für den Verschlüsselungsvorgang generierte Schlüssel von *IDEA* werden mit dem öffentlichen Schlüssel des Mitlieds verschlüsselt. Für die Generierung der Schlüsselpaare des asymmetrischen Verfahrens, mit denen der *IDEA*-Schlüssel und der Wahlschein verschlüsselt werden, verwenden wir den *RSA*-Algorithmus, der erste vollständige und heute am weitesten verbreitete asymmetrische Kryptographiealgortihmus.[28] Das Schlüsselpaar wird bei *RSA* mit Hilfe des Produkts zweier (sehr großer) Primzahlen gebildet. Die Sicherheit des Algorithmus beruht darauf, daß bisher kein einfaches Verfahren gefunden werden konnte, Zahlen dieser Größenordnung zu faktorisieren. *IDEA* ist durch ein europäisches Patent geschützt;[29] seine Verwendung für nicht-kommerzielle Zwecke ist aber freigestellt. *RSA* ist lediglich in den USA patentiert.[30]

3.3 Softwarekomponenten

Um das Verfahren wie dargelegt durchführen zu können, sind drei Softwarekomponenten erforderlich:

- Das Modul für die Generierung der Schlüsselpaare („Zertifizierungsstelle“; „Wahlamt“),
- die eigentliche Wahlsoftware, ein WWW-Client, realisiert als Java™-Applet: „Wahlkabine“ mit „Wahlschein“ (der öffentliche Schlüssel) und „Wahlzettel“ (das Java™-Formular) und
- das Modul zur Entschlüsselung der Wahlscheine, das auch Auswertungsausgaben übernehmen kann („Wahlurne“ und „Wahlleiter“).

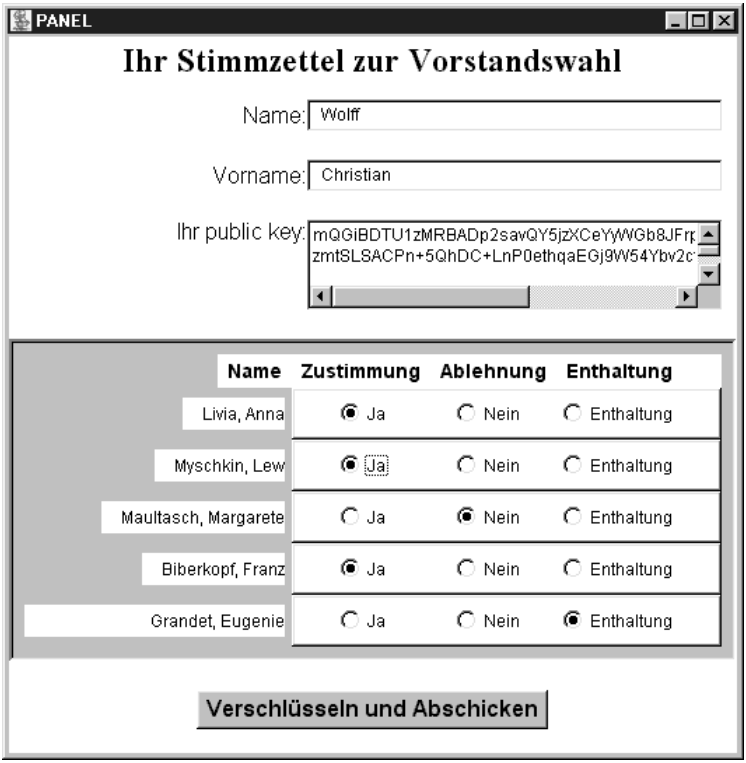


Abb. 1: Prototypische Realisierung von Wahlschein und Stimmzette

Bei der Implementierung stehen die folgenden Aspekte im Mittelpunkt:

1) Der Einsatz der Programmiersprache Java™ gewährleistet plattformübergreifende Verfügbarkeit.[31]
2) Sichere Kryptographie verlangt – auch um das Vertrauen der Benutzer zu gewinnen – nach einem Höchstmaß an Transparenz. Deshalb kommen nur kryptographische Verfahren zum Einsatz, deren Algorithmen öffentlich sind, und die bereits intensiver Kryptanalyse unterzogen wurden.
3) Alle Prozesse lassen sich weitgehend automatisieren, um den Arbeitsaufwand (des Wählers wie des Wahlleiters) im Vergleich zum traditionellen Wahlverfahren zu minimieren. Dabei wird zur Schlüsselgenerierung bzw.

-authentisierung auch auf eine Mitgliedsdatenbank zurückgegriffen.

4) Für den vielfach im Umgang mit kryptographischen Verfahren unerfahrenen Benutzer steht ein einfaches Interface zur Verfügung. Die wesentliche Hürde besteht aus dem Transfer der Schlüsseldaten aus dem e-mail-Viewer des Wählers in das entsprechende Textfeld des Wahlscheins.
5) Die modulare Struktur soll die Generalisierbarkeit der Anwendung sicherstellen. Dabei ist an eine automatische Generierung passender Wahlformulare für unterschiedliche Typen von Wahlen, Umfragen oder Erhebungen zu denken. Aufgrund der gegebenen formalen Beschreibbarkeit der abzufragenden Daten erscheint dies problemlos möglich.

Abb. 1 zeigt den Prototyp des Webclients mit einem hypothetischen Wahlzettel. Er sieht die Zustimmung, Ablehnung und Enthaltung als typische Wahlmöglichkeiten vor. Darüber hinaus kann im einzelnen auch *keine Stimmabgabe* erfolgen, wenn kein Radio Button angekreuzt ist.

4 Zur praktischen Umsetzung

Die entscheidenden Hindernisse auf dem Weg zum Einsatz kryptographiebasierter Kommunikationsverfahren liegen weniger auf dem Feld technischer Realisierbarkeit als vielmehr im Bereich der Organisation und der Logistik sowie der Akzeptanz durch den Benutzer. Im konkreten Einsatzgebiet – elektronische Vorstandswahlen der GLDV – bedeutet dies:

1) Das Verfahren wird den Mitgliedern frühzeitig vorgestellt und auf Mitgliederversammlungen diskutiert, ggf. modifiziert.
2) Vor dem eigentlichen Einsatz erfolgen Testläufe mit hypothetischen Daten, um technische und ergonomische Schwachstellen erkennen und beseitigen zu können.
3) Es wird nur parallel und in Ergänzung zum traditionellen papierbasierten Wahlverfahren eingesetzt; jedes Mitglied hat also die Möglichkeit, das für ihn einfachere Wahlverfahren zu verwenden.

5 Fazit & Ausblick

"Good design starts with a threat model: what the system is to protect, from whom, and for how long."[32] Dieser Devise von Bruce SCHNEIER folgend sei gefragt, welche Sicherheitslücken und Angriffspunkte das vorgeschlagene Verfahren bietet: Der entscheidende Schwachpunkt ist wohl die einfache Versendung der öffentlichen Mitgliedsschlüssel per e-mail, da diese im Internet abgefangen werden könnten. Zwar ist ein mehrfaches Wirksamwerden einer Stimmabgabe mit demselben Schlüssel ausgeschlossen, aber es könnte ein Unbefugter dem berechtigten Mitglied die Stimme vorenthalten, indem er ohne Berechtigung das Stimmrecht ausübt und so das Wahlergebnis verfälscht. Im Rahmen des gewählten Testszenarios halten wir diesen Fall aber für hinreichend unwahrscheinlich, um auf weitere, das Verfahren verkomplizierende, Sicherungsmaßnahmen verzichten zu können.

Mittelfristig ist das Verfahren auf nach dem SigG von einer amtlichen Zertifizierungsstelle beglaubigte digitale Signaturen umzustellen. Auf diese Variante haben wir zunächst verzichtet, da wir davon ausgehen, daß nur sehr wenige Wahlberechtigte über eine derartige Signatur verfügen.

Die elektronische Wahl ist der am einfachsten zu formalisierende und automatisierende Kommunikationsprozeß im Verbandsleben. Wollte man z. B. eine Mitgliederversammlung elektronisch durchführen, so wäre einerseits eine wesentlich komplexere kommunikative Struktur zu modellieren (Einberufung, Tagesordnung, Eröffnung, Genehmigung der Tagesordnung, Anträge aus der Mitte der Mitgliederversammlung, Bericht und Diskussion, Abstimmung und Wahlen), was nur durch die Verwendung heterogener Software-Systeme denkbar erschiene (e-mail-Verteiler, Diskussionsforen, *electronic vote*, Videokonferenz). Andererseits ist es von vornherein fraglich, ob das ineinandergreifende Geflecht unterschiedlicher Rechte der Verbandsorgane und Mitglieder, das in einer Mitgliederversammlung zum Tragen kommt (Teilnahmerecht, Rederecht, Auskunftsrecht, Antragsrecht, Stimmrecht)[33] sich überhaupt mit elektronischen Mitteln modellieren läßt. Es erscheint daher sinnvoller, neben elektronischen Wahlen weitere rechnergestützte Kommunikationsformen als neue Vereinsorgane in die Satzung aufzunehmen. Auf diesem Weg könnten z.B. der Mitgliederversammlung vorbehaltene Entscheidungen auch auf elektronisch durchgeführte Diskussionen und Abstimmungen übertragen werden, ohne daß dies einer „virtuellen Mitgliederversammlung" gleichkäme.

Anmerkungen

[1]Vgl. RANNENBERG, MÜLLER & PFITZMANN 1997:22f; SUN MICROSYSTEMS 1997B:1.

[2]Zur Einführung in kryptographische Verfahren und die wichtigsten Algorithmen vgl. WOBST 1997, bes. 105ff, 136ff.

[3]Vgl. IANNAMICO 1997; SCHNEIER 1996:664ff; WOBST 1997:275ff.

[4]REICHERT & VAN LOOK 1995:Rdnr. 259.

[5]REICHERT & VAN LOOK 1995:Rdnr. 260.

[6]BGHZ 47, 172 (177) = NJW 1967, 1268 (1270); BGHZ 105, 306 (313 f.) = NJW 1989, 1724 (1725); zustimmend auch die herrschende Meinung in der Literatur: MÜNCHKOMM/*REUTER* § 25 Rdnr. 3; STAUDINGER/*WEICK* § 25 Rdnr. 3; REICHERT & VAN LOOK 1995:Rdnr. 262 mit weiteren Nachweisen sowie Rdnr. 315 ff.

[7]REICHERT & VAN LOOK 1995:Rdnr. 279.

[8]BGHZ 105, 306 (314) = NJW 1989, 1724 (1725); REICHERT & VAN LOOK 1995:Rdnr. 279; a. M.: MÜNCHKOMM/*REUTER* § 25 Rdnr. 6, der die Sicherung der Integration der Vereinsverfassung als Hauptzweck ansieht.

[9]REICHERT & VAN LOOK 1995:Rdnr. 263.

[10]STAUDINGER/*WEICK* § 25 Rdnr. 4.

[11]REICHERT & VAN LOOK 1995:Rdnr. 319.

[12]GEIS 1997:3000.

[13]GEIS 1997:3002.

[14]Nur ausnahmsweise sind zwingend Formvorschriften zu beachten, etwa gem. § 37 Abs. 1 BGB, nach dem der Antrag einer Minderheit auf Einberufung der Mitgliederversammlung „schriftlich unter Angabe des Zweckes und der Gründe“ zu stellen ist. Ebenso sind die Formvorschriften über die Bekanntmachung der Auflösung des Vereins oder der Entziehung der Rechtsfähigkeit (§ 50 BGB) zwingendes Recht.

[15]WALDNER & RÖSELER 1994:94, Rdnr. 130.

[16]REICHERT & VAN LOOK 1995:Rdnr. 301; STAUDINGER/*WEICK* § 25 Rdnr. 16.

[17]BGH NJW 1994, 51 [52].

[18]STAUDINGER/*WEICK* § 25 Rdnr. 16.

[19]Vgl. BGH NJW 1954, 953; BGHZ 47, 172 (177) = NJW 1967, 1268 (1270); BGHZ 47, 381 [386].

[20]MÜNCHKOMM/*REUTER* § 25 Rdnr. 3; vgl. auch REICHERT & VAN LOOK 1995:Rdnr. 263, 296.

[21]MÜNCHKOMM/*REUTER* § 25 Rdnr. 4 mit einer Auflistung der dazu ergangenen Rechtsprechung sowie der Ansicht in der Literatur; vgl. auch REICHERT & VAN LOOK 1995:Rdnr. 315.

[22]REICHERT & VAN LOOK 1995:Rdnr. 1227: „Im Regelfall wird der Vorstand in der Mitgliederversammlung gewählt.“

[23]Eine Zertifizierungsstelle im Sinne des SigG ist eine Behörde oder ein von einer Behörde beauftragtes Unternehmen, daß Signaturschlüssel und digitale Zertifikate generiert, ausgibt und verwaltet, d. h. zur Verfikation in einer Datenbank bereithält, vgl. BIESER 1997:402ff.

[24]Vgl. SUN MICROSYSTEMS 1997A.

[25]Dies gilt aber nicht für kryptographische Verfahren, die lediglich der Authentisierung dienen, also etwa digitale Signaturen, da sie die eigentliche Nachricht nicht verschlüsseln. Deshalb sind

die Algorithmen für digitale Signaturen frei verfügbar. Vgl. SCHNEIER 1996:691ff; WOBST 1996:275ff, 307ff.

[26]Vgl. PLATZER 1998.

[27]Vgl. SCHNEIER 1996:370ff; WOBST 1997:182ff.

[28]Der RSA-Algorithmus ist *vollständig*, da er sowohl für Verschlüsselung als auch für digitale Signaturen geeignet ist; RSA ist nach seinen Entwicklern RIVEST, SHAMIR und ADLEMAN benannt, vgl. SCHNEIER 1996:531ff; WOBST 1997:143ff.

[29]*IDEA*: Europ. Patent Nº 0482154 vom 30.6.1991.

[30]Vgl. WOBST 1997:154.

[31]Vgl. ARNOLD & GOSLING 1997.

[32]SCHNEIER 1997:3.

[33]Vgl. REICHERT & VAN LOOK 1995:Rdnr. 869ff, 880ff, 885ff, 890ff, 895ff.

Literatur

ARNOLD, Ken; GOSLING, James (1997[2]). The Java™ Programming Language. Reading/MA et al.: Addison-Wesley.

BIESER, Wendelin (1997). „Begründung und Überlegung zum Signaturgesetz." In: MÜLLER & PFITZMANN (1997), 399 – 410.

GARFINKEL, Simson; SPAFFORD, Gene (1997). „Cryptography and the Web." In: World Wide Web Journal 2(4) (1997), 113 – 126.

GEIS, Ivo (1997). „Die Digitale Signatur." In: Neue Juristische Wochenschrift 1997, 3000 – 3004.

IANNAMICO, Mike (1997). Pretty Good Privacy™. PGP for Personal Privacy, Version 5.0 for Windows® 95, Windows NT. User's Guide. San Mateo/CA: Pretty Good Privacy, Inc.

KHARE, Rohit; RIFKIN, Adam (1997). „Weaving a Web of Trust." In: World Wide Web Journal 2(4) (1997), 77 – 112.

MÜNCHKOMM/*REUTER*. Münchener Kommentar zum Bürgerlichen Gesetzbuch, Bd. 1, Allgemeiner Teil, 3. Auflage, München: Beck 1993, zit. als MÜNCHKOMM/*REUTER*.

MÜLLER, Günter; PFITZMANN, Andreas (edd.) (1997). Mehrseitige Sicherheit in der Kommunikationstechnik. Verfahren, Komponenten, Integration. Bonn et al.: Addison-Wesley.

PLATZER, Wolfgang (1998). The IAIK Java Cryptography Extension. TU Graz, Institute for Applied Information Processing and Communications. März 1998. http://jcewww.iaik.tu-graz.ac.at/IAIK_JCE/jce.htm . (22. Mai 1998).

RANNENBERG, Kai; MÜLLER, Günter; PFITZMANN, Andreas (1997). Sicherheit, insbesondere mehrseitige IT-Sicherheit. In: MÜLLER & PFITZMANN (1997), 2129.

REICHERT, Bernhard; VAN LOOK, Frank (1995). Handbuch des Vereins- und Verbandsrechts. Neuwied et al.: Luchterhand.

SCHNEIER, Bruce (1996). Angewandte Kryptographie. Bonn et al.: Addison-Wesley.

SCHNEIER, Bruce (1997). Why Cryptography Is Harder Than it Looks. Technical Report, Counterpane Systems Inc, Minneapolis.

SCHUSTER, Rolf; FÄRBER, Johannes; EBERL, Markus (1997). Digital Cash. Zahlungssysteme im Internet. Berlin et al.: Springer.

STAUDINGER/*WEICK*. J. v. STAUDINGERS Kommentar zum Bürgerlichen Gesetzbuch mit Einführungsgesetz und Nebengesetzen, 13. Bearbeitung, Berlin: Sellier-de Gruyter 1995.

SUN MICROSYSTEMS Inc. (1997A). Java Security Architecture. October 1997.

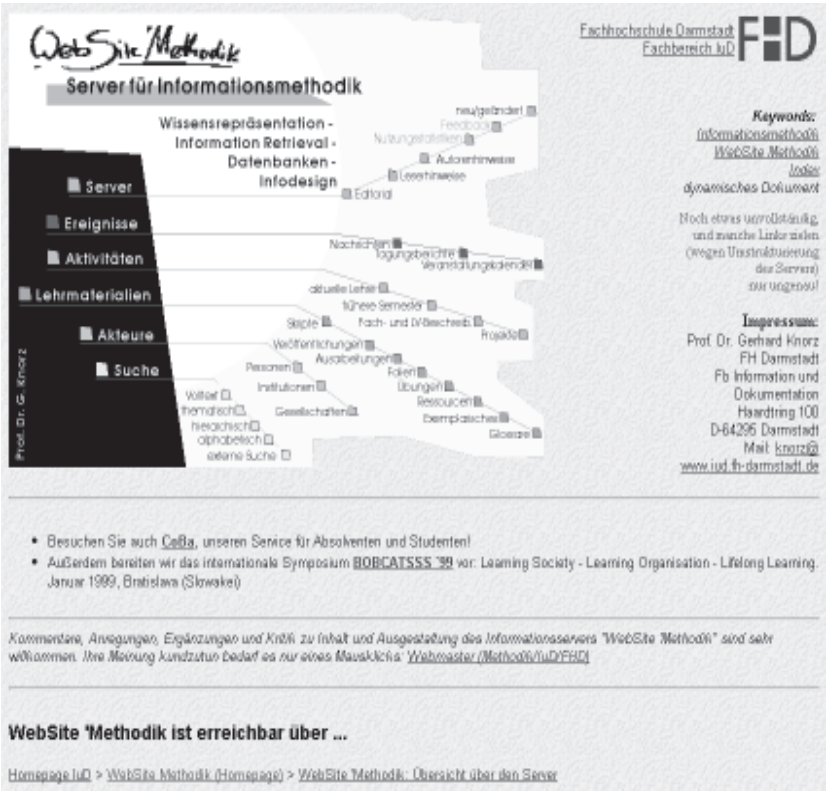
SUN MICROSYSTEMS Inc. (1997B). Secure Computing with Java: Now and the Future. A White Paper. 1997. http://www.javasoft.com/marketing/collateral/security.html . (22. Mai 1997).

WALDNER, Wolfram; RÖSELER, Diana (1994[15]). Der eingetragene Verein. München: Beck.

WOBST, Reinhard (1997). Abenteuer Kryptographie. Methoden, Risiken und Nutzen der Datenverschlüsselung. Bonn et al.: Addison-Wesley.

„WebSite 'Methodik“ – eine gemeinsame Informationsressource für Hochschullehrer und Studierende

Gerhard Knorz



Die Übersichtsseite zu* WebSite 'Methodik*, dem Informationsserver für das Fach Informationsmethodik am Fachbereich Information und Dokumentation der Fachhochschule Darmstadt. Die Inhalte sind in 5 Rubriken gegliedert, dazu kommen verschiedene Formen der Suche und Übersichten.

1 Ein Blick über die Schulter ...

Frau Sabine Studor, Studentin des 2. Semesters im Studiengang Information und Dokumentation (IuD) der Fachhochschule Darmstadt, hat ihre Unterlagen für die Gruppenübung zusammengepackt und den letzten Schluck Kaffee geleert. Nun drängt sie ihre Kommilitonen, denn sie muß noch aus dem Gruppenarbeitsraum zu dem PC-Pool wechseln: Sie will die Öffnungszeit nutzen, um sich auf die morgige Lehrveranstaltung „Informationsmethodik II" vorzubereiten: Hat sie alle wichtigen Folien vorliegen (sie druckt sich diese eigentlich immer aus – für den dicken LV-Ordner)? Gibt es morgen eine Übung, die sie parat haben muß? Und überhaupt: Um was geht es morgen eigentlich? Der freie Zugang zu PCs und Internet ist für Studierende am Fachbereich IuD durchaus noch ein Problem, nachdem die Einrichtung eines studentenverwalteten Lernraumes sich länger hinzieht als gedacht. Aber nun sitzt Frau Studor im sogenannten „Projekt-Labor" und holt sich per Bookmark die Übersichtsseite (http://www.iud.fh-darmstadt.de) für *Web-Site 'Methodik* auf den Bildschirm (Titelgrafik).

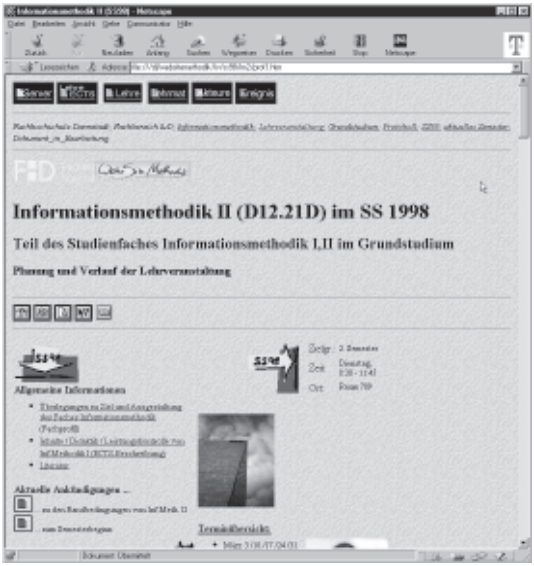
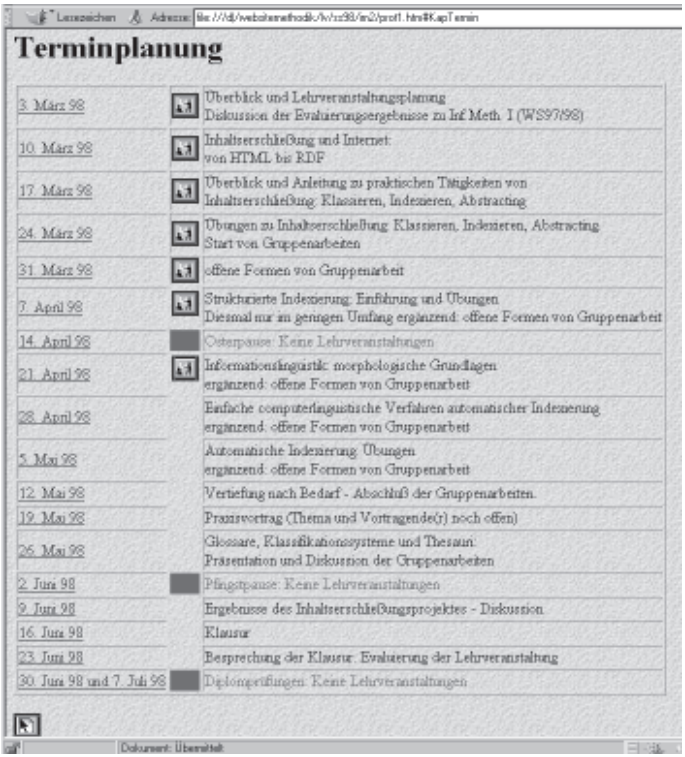


Abb. 1: *Die Einstiegsseite zu den Lehrveranstaltungen faßt alle Informationen zusammen (bzw. verweist auf sie), wie sie üblicherweise in kommentierten Vorlesungsverzeichnissen stehen, ergänzt um aktuelle Ankündigungen, Arbeitsergebnisse und die detaillierte Veranstaltungsplanung (a priori), wobei letztere nach Überarbeitung zur Veranstaltungsdokumentation wird (a posteriori).*

Sabine Studor hält sich nicht weiter mit dem ausdifferenzierten Angebot auf: Der Unterpunkt „Aktuelle Lehre“ der Rubrik „Aktivitäten“ bringt sie zu einer weiteren „Übersichtskarte“, in der Sie „ihre“ Lehrveranstaltung auswählt: „Informationsmethodik II“ (Abb. 1).

Auch hier hält sich Frau Studor nicht auf: Seitdem sie sich in die (geschützte) Mailingliste zu Informationsmethodik II eingetragen hat, sind ihr die „Ankündigungen“ (sofern es aktuelle gibt) bereits bekannt und sie folgt direkt dem Link zur Terminübersicht (Abb. 2).



Terminplanung

3. März 98	Überblick und Lehrveranstaltungsplanung Diskussion der Evaluierungsergebnisse zu Inf.Meth. I (WS97/98)
10. März 98	Inhaltserschließung und Internet: von HTML bis RDF
17. März 98	Überblick und Anleitung zu praktischen Tätigkeiten von Inhaltserschließung: Klassieren, Indexieren, Abstracting
24. März 98	Übungen zu Inhaltserschließung: Klassieren, Indexieren, Abstracting Start von Gruppenarbeiten
31. März 98	offene Formen von Gruppenarbeit
7. April 98	Strukturierte Indexierung: Einführung und Übungen Diesmal nur in geringen Umfang ergänzend: offene Formen von Gruppenarbeit
14. April 98	Osterpause: Keine Lehrveranstaltungen
21. April 98	Informationslinguistik: morphologische Grundlagen ergänzend: offene Formen von Gruppenarbeit
28. April 98	Einfache computerlinguistische Verfahren automatischer Indexierung ergänzend: offene Formen von Gruppenarbeit
5. Mai 98	Automatische Indexierung: Übungen ergänzend: offene Formen von Gruppenarbeit
12. Mai 98	Vertiefung nach Bedarf - Abschluß der Gruppenarbeiten.
19. Mai 98	Praxisvortrag (Thema und Vortragende(r) noch offen)
26. Mai 98	Glossare, Klassifikationssysteme und Thesauri: Präsentation und Diskussion der Gruppenarbeiten
2. Juni 98	Pfingstpause: Keine Lehrveranstaltungen
9. Juni 98	Ergebnisse des Inhaltserschließungsprojektes - Diskussion
16. Juni 98	Klausur
23. Juni 98	Besprechung der Klausur: Evaluierung der Lehrveranstaltung
30. Juni 98 und 7. Juli 98	Diplomprüfungen: Keine Lehrveranstaltungen

Abb. 2: *Terminplanung zur Lehrveranstaltung*

Ein Blick zeigt, daß nach der Osterpause (fast) die Hälfte des Semesters bereits bewältigt ist und daß sich am Zeitplan offensichtlich nichts aktuell geändert hat.

Ein Klick auf den „21. April“ führt nun zur Planung des anstehenden Veranstaltungstermins (Abb. 3).

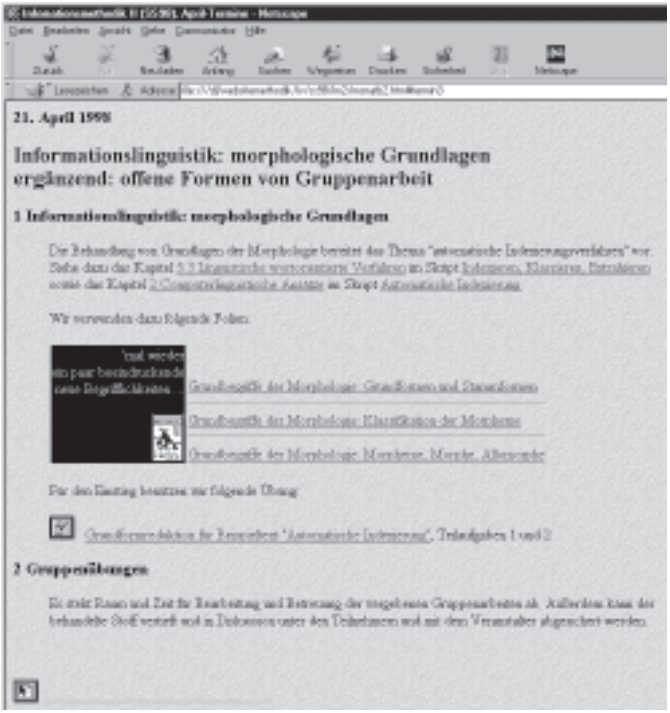


Abb. 3: *Planung eines Lehrveranstaltungstermins mit allen verwendeten Unterlagen (Skripte, Folien, Übungen). Der Umfang dieser Darstellungen schwankt beträchlich: Der vorliegende Fall zählt zu den eher kürzeren Beispielen. Die Lehrveranstaltungsplanung wird nach Überarbeitung zur Dokumentation des tatsächlichen Lehrveranstaltungsverlaufs.*

„Grundbegriffe der Morphologie“, das ist also das Thema. Frau Studor atmet auf: Diese Folien (es sind insgesamt 14 einzelne Folien, zusammengefaßt in 3 „Foliensequenzen“) und auch die Übung hat sie zu einem früheren Zeitpunkt bereits heruntergeladen und ausgedruckt – es ist nichts Neues hinzugekommen. Sie läßt sich dennoch dazu verleiten, einem der Links auf Folien zu folgen, weil in ihren Ohren die Begrifflichkeit gar so befremdlich klingt: Morphe, Allomorphe, ...? (Abb. 4).

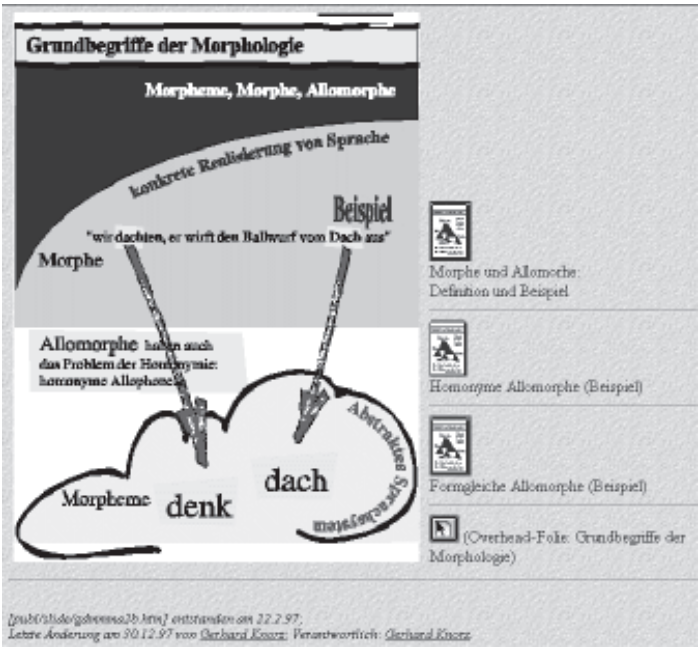


Abb. 4: *Beispiel einer Folie, als zweite einer Foliensequenz. Die Darstellungen auf Folien sind bewußt plakativ mit groben Formen und großflächigen Farben gestaltet. Die physische Realisierung der Foliensequenzen benutzt einzelne Folien, die übereinandergelegt und ausgetauscht werden. Im Rechner liegen eine qualitativ hochwertige Repräsentation im Vektorformat vor (nicht öffentlich zugänglich), dazu eine Bildschirmauflösung (400 * 530 Pixel, 256 Farben) und eine gepackte Version zum Download (600*800 Pixel, 256 Farben).*

Frau Studor erweist sich als interessierte Studentin: Sie hat mit dem Link auf die Folie automatisch die Rubrik gewechselt (von „Aktivitäten“ zu „Lehrmaterial“) und wechselt nun anschließend mit einem weiteren Mausklick bewußt die Hierarchieebene. Damit kommt sie auf eine thematisch gegliederte Übersicht aller Folien: Gegenwärtig (April 1998) werden 326 Einzelfolien, zusammengefaßt in 143 Dokumenten (die entweder eine einzelne Folie oder aber eine Foliensequenz darstellen), untergliedert nach ca. 50 Themen unter knapp 15 Hauptüberschriften. Eine dieser Hauptüberschriften ist „Informationslinguistik“ mit u. a. dem Thema „Morphologie“. Hier findet sie die vollständige Auslistung aller einschlägigen Folien, jeweils mit einer kategorisierten Beschreibung (Abb. 5).

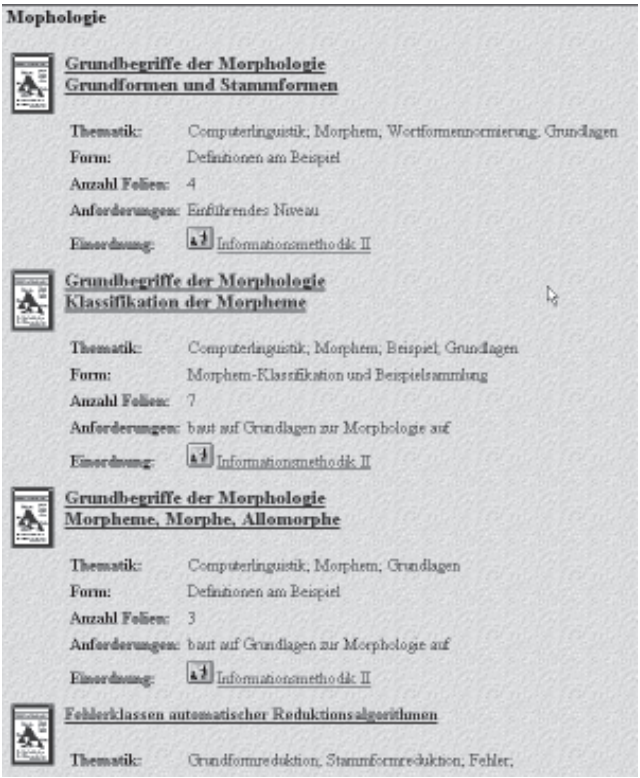
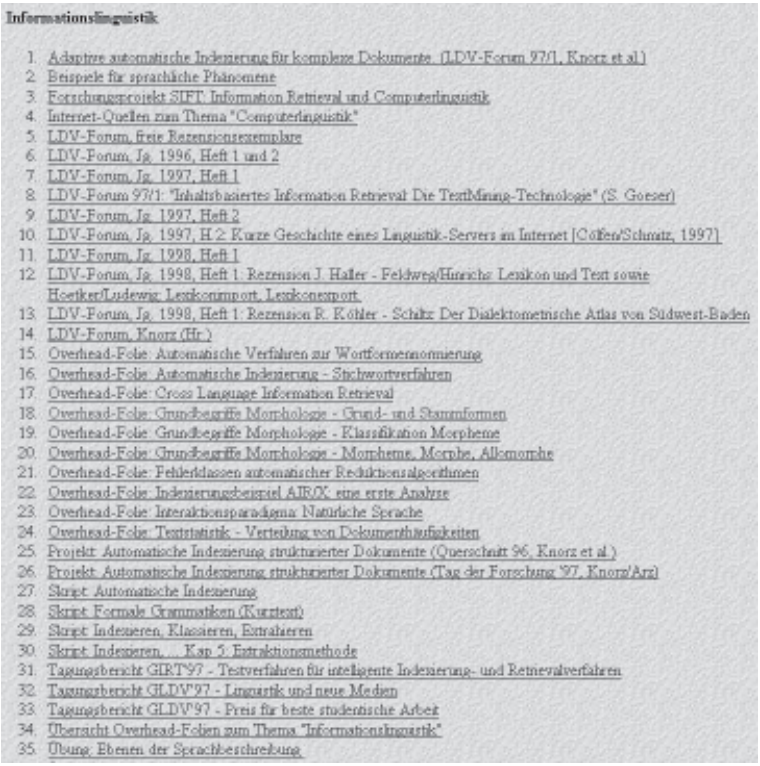


Abb. 5: *Übersicht über Folien zum Thema „Morphologie". Alphabetisch geordnet, mit kategorisierter Beschreibung.*

Jedes Dokument in *WebSite 'Methodik* ist manuell indexiert, d.h. es sind ihm mehrere Schlagwörter (Deskriptoren) aus einem kontrollierten Vokabular mit ca. 100 Begriffen zugeordnet. Diese Indexierung hat ihren festen Platz im Header jedes Dokumentes unterhalb der Rubrikleiste. Siehe dazu Abb. 1: Bei diesem Dokument, „Informationsmethodik II im SS 1998", besteht die Indexierung aus den Deskriptoren *Fachhochschule Darmstadt, Fachbereich IuD, Informationsmethodik, Grundstudium, Protokoll, SS98, aktuelles Semester, Dokument_in_Bearbeitung*. Die Folie „Grundbegriffe der Morphologie ..." (Ausschnitt in Abb. 4) ist u. a. mit dem Deskriptor *Informationslinguistik* indexiert. Über diese

Indexierung gelangt man in das „Schlagwortregister“ und findet dort unmittelbar die Zusammenstellung aller Dokumente in *WebSite 'Methodik*, die diesem Thema zuzurechnen sind (Abb. 7). Frau Studor findet dort 41 Dokumente, darunter Beispiele, mehrere Beiträge des LDV-Forum, 10 Folien(-sequenzen), Skripte, Tagungsberichte und u. a. auch 6 Übungen. Sie kann also auf einfache Weise ausgehend von einem konkreten Dokument weitere Dokumente (auch anderen Typs) desselben Themas finden – unabhängig von einzelnen Lehrveranstaltungen.



Informationslinguistik

1. Adaptive automatische Indexierung für komplexe Dokumente (LDV-Forum 97/1, Knorz et al.)
2. Beispiele für sprachliche Phänomene
3. Forschungsprojekt SIFE: Information Retrieval und Computerlinguistik
4. Internet-Quellen zum Thema "Computerlinguistik"
5. LDV-Forum, freie Rezensionsexemplare
6. LDV-Forum, Jg. 1996, Heft 1 und 2
7. LDV-Forum, Jg. 1997, Heft 1
8. LDV-Forum 97/1: "Inhaltsbasiertes Information Retrieval: Die TextMining-Technologie" (S. Goeser)
9. LDV-Forum, Jg. 1997, Heft 2
10. LDV-Forum, Jg. 1997, H.2: Kurze Geschichte eines Linguistik-Servers im Internet [Cölfen/Schmitz, 1997]
11. LDV-Forum, Jg. 1998, Heft 1
12. LDV-Forum, Jg. 1998, Heft 1: Rezension J. Haller - Feldweg/Hinrichs: Lexikon und Text sowie Hoetker/Ludewig: Lexikonimport, Lexikonexport
13. LDV-Forum, Jg. 1998, Heft 1: Rezension R. Köhler - Schiltz: Der Dialektometrische Atlas von Südwest-Baden
14. LDV-Forum, Knorz (Hr.)
15. Overhead-Folie: Automatische Verfahren zur Wortformennormierung
16. Overhead-Folie: Automatische Indexierung - Stichwortverfahren
17. Overhead-Folie: Cross Language Information Retrieval
18. Overhead-Folie: Grundbegriffe Morphologie - Grund- und Stammformen
19. Overhead-Folie: Grundbegriffe Morphologie - Klassifikation Morpheme
20. Overhead-Folie: Grundbegriffe Morphologie - Morpheme, Morphe, Allomorphe
21. Overhead-Folie: Fehlerklassen automatischer Reduktionsalgorithmen
22. Overhead-Folie: Indexierungsbeispiel AIR/X: eine erste Analyse
23. Overhead-Folie: Interaktionsparadigma: Natürliche Sprache
24. Overhead-Folie: Textstatistik - Verteilung von Dokumenthäufigkeiten
25. Projekt: Automatische Indexierung strukturierter Dokumente (Querschnitt 96, Knorz et al.)
26. Projekt: Automatische Indexierung strukturierter Dokumente (Tag der Forschung '97, Knorz/Artz)
27. Skript: Automatische Indexierung
28. Skript: Formale Grammatiken (Kurztext)
29. Skript: Indexieren, Klassieren, Extrahieren
30. Skript: Indexieren, ... Kap. 5: Extraktionsmethode
31. Tagungsbericht GIRT'97 - Testverfahren für intelligente Indexierungs- und Retrievalverfahren
32. Tagungsbericht GLDV'97 - Linguistik und neue Medien
33. Tagungsbericht GLDV'97 - Preis für beste studentische Arbeit
34. Übersicht Overhead-Folien zum Thema "Informationslinguistik"
35. Übung: Ebenen der Sprachbeschreibung

Abb. 7: *Übersicht über Dokumente zum Thema Informationslinguistik (35 von insgesamt 41 Dokumenten). Diese Übersicht ist von jedem Dokument aus erreichbar: entweder direkt über die Deskriptoren der Indexierung (hier also über den Deskriptor „Informationslinguistik“) oder aber über das Such-Icon, das auf verschiedene Arten der Suche führt, u. a. auf dieses Register.*

2 *WebSite 'Methodik* – Entwicklung und Rolle für das Lehrkonzept

2.1 Rahmendaten

Die Geschichte von *WebSite 'Methodik* geht zurück bis zum Sommersemester 1995, als ein Informationsserver für den Fachbereich IuD in technischer und inhaltlicher Hinsicht im Rahmen eines studentischen Projektes konzipiert und realisiert wurde. Dieser Server stellte für jedes Fach und jeden Hochschullehrer Basisinformationen bereit, die im Falle des Faches Informationsmethodik unmittelbar weiter ausgebaut wurden. Innerhalb kurzer Zeit war die Balance zwischen dem Informationsbestand auf Fachbereichs- und Fachebene einerseits und dem Informationsangebot für das spezifische Fach Informationsmethodik andererseits soweit verlorengegangen, so daß eine klare Trennung zwischen beiden Informationswelten die sinnvolle Konsequenz war. Im Sommersemester 1996 wurde dann die inhaltliche und strukturelle Konzeption von *WebSite 'Methodik* als dem Internet-Angebot des Faches Informationsmethodik entwickelt und die Umstrukturierung der vorhandenen Ressourcen vollzogen. Der Fachbereichsserver ist seitdem ein Informationsangebot mit (im wesentlichen) semesterweiser Aktualisierung unter der Verantwortung des Dekans. *WebSite 'Methodik* wird im wesentlichen tagesaktuell gehalten und ist als selbständiger Service über die Seiten des Fachbereichs erreichbar. Seit Wintersemester dokumentiert der Server lükkenlos alle im Bereich der Lehre relevanten Aktivitäten (auf der Ebene der einzelnen Termine und Inhalte) und etliches darüber hinaus.

WebSite 'Methodik besteht mittlerweile aus ca. 420 Dokumenten, die auf ca. 2 500 Dateien in ca. 200 Verzeichnissen basieren (85 MB). Beispielsweise bilden alle Arbeitsergebnisse studentischer Arbeiten im Rahmen einer Lehrveranstaltung zusammen mit allen anderen zugeordneten Informationen zunächst ein einziges logisches Dokument, das auf mehrere Dateien verteilt ist. Die Änderungsrate beträgt, bedingt durch die notwendigen Nacharbeitungen der Lehrveranstaltungsplanung, etwa 20 bis 30 Dateien/Woche. Das Volumen hat sich im Laufe des vergangenen Jahres in etwa verdoppelt.

2.2 Anspruch und Realität

Entwurf und Ausgestaltung von *WebSite 'Methodik* wird von folgenden Zielen geleitet:

- **Dienstleistung fürs Studieren** ist das selbstverständliche erste Ziel des Internet-basierten Service:
  - Bereitstellen von *Unterlagen für laufende Lehrveranstaltungen und deren Vorbereitung*, insbesondere auch Verweise auf Demonstrationssysteme und Werkzeuge wie z. B. Fachvokabularien (Thesauri)
  - Unterstützung von *Nachbereitung, Vertiefung und Prüfungsvorbereitung*
  - *Orientierung* über Lehrveranstaltungs- und Semestergrenzen hinweg
  - *Unterstützung der Kommunikation* zwischen Veranstalter und Gruppen von Studierenden, sowie zwischen den Teilnehmern von Veranstaltungen (Gruppenarbeiten!)
- **Stärkung der Autonomie** der Studierenden. Die angebotenen Informations- und Kommunikationsmöglichkeiten sollen unabhängiger von Raum und Zeit machen und eine interessensgeleitete Vertiefung des Informationsmethodik-Stoffes unterstützen. Insbesondere soll es einladen dazu, die weltweiten Ressourcen des Internet aktiv in die Beschäftigung mit den Lehrveranstaltungsthemen einzubeziehen.
- **Transparenz** soll in verschiedener Hinsicht erreicht werden:
  - Jedem Kollegen, jedem Studierenden und jedem Interessierten soll erkennbar sein, was *konkret* in welcher 'Methodik-Veranstaltung behandelt wird.
  - Ergebnisse der Evaluierung von Lehrveranstaltungen werden systematisch durchgeführt und „unzensiert“ zum öffentlichen Bestandteil der Lehrveranstaltungsdokumentation gemacht.
  - *WebSite 'Methodik* steht für jeden Nutzer nur wenige Hyperlinks von den einschlägigen WWW-Servern anderer informationswissenschaftlicher Ausbildungseinrichtungen entfernt. Vergleiche sind kein Mißbrauch!
- **Eine aktive Rolle der Studierenden** im Lehr- und Lernprozess wird primär durch spezifische Lehrformen in Informationsmethodik „erzwungen“, aber von *WebSite 'Methodik* immerhin dahingehend gefördert, daß studentische Ausarbeitungen und Arbeitsergebnisse Teil des Informationsservers werden.

- **Freude am Lernen** soll durch eine ästhetische und inhaltliche Qualität von *WebSite 'Methodik* gefördert werden.
- **Als gemeinsame Informationsressource** für den Lehrenden *und* die Studierenden soll *WebSite 'Methodik* zur Verfügung stehen: Sie ist die (dokumentenorientierte) Wissensbasis, in der der Fachvertreter seine Informationen strukturiert ablegt und aus der heraus die Lehrveranstaltungen, Kurse, Vorträge etc. vorbereitet und dokumentiert werden.

WebSite 'Methodik hat den papiergebundenen Informationsfluß zwischen Veranstalter und Studierenden auf nahezu Null reduziert, was nicht ausschließt, daß Studierende ihrerseits die elektronische Darstellung doch wieder in eine Printform transformieren. Der bisher vertretene Ansatz, Lehre im Fach Informationsmethodik als ein Distance Learning im Dual Mode zu vertreten, also Formen des zeit- und ortsungebundenen Lernens als Add-on zu Präsensveranstaltungen anzubieten, wird auf absehbare Zeit bestehen bleiben. Präsenzgebundene Formen der Lehre wie Gruppenarbeiten, differenzierte Aufgaben und Rollen der TeilnehmerInnen im Lehr- und Lernbetrieb haben einen hohen Stellenwert neben den elektronischen Vermittlungsformen. Bereits bei Studierenden im ersten Semester, in dem die freien Zugangsvoraussetzungen zum Netz unter den bestehenden Gegebenheiten des Fachbereichs gegenüber höheren Semestern weitaus reduzierter sind, zeigt sich, daß *WebSite 'Methodik* als Angebot von einigen positiv hervorgehoben wird, wenngleich die Furcht noch überwiegt, Wichtiges könne an einem vorbeigehen: In einer aktuellen ausführlichen Evaluierung von Informationsmethodik I wurden abschließende freie Kommentare z. T. auch mit Bezug auf *WebSite 'Methodik* abgegeben:

Positives Feedback

- Der aktuelle Bezug zu modernen Medien;
- Einsatz neuer Medien (HTML);
- die vielen Übungen, Gruppenarbeiten, Folien zur LV sind per Internet abrufbar und damit bessere Information;
- gute Strukturierung/Gliederung des Stoffes;
- selbständiges Arbeiten mit Thesauri und Klassifikations-Systemen (im Web verfügbar).

Kritische Kommentare

- Unterrichtsmaterial nur im Internet: Skript nur im Internet;
- daß man das Skript aus dem Internet abrufen muß;
- daß Skripte nicht als Print vorhanden sind, sondern daß ein Web-Zugang Voraussetzung ist, um LV-Materialien zu erlangen.
- Zu häufiger Verweis auf *WebSite 'Methodik*, auf die ich noch keinen Zugriff habe, deshalb Zugriff auf Lehrinhalte nur mit erheblichem Mehraufwand möglich.

Die vollständigen Ergebnisse der Lehrveranstaltungs-Evaluierung vom Wintersemester 1997/98 sind unter http://www.iud.fh-darmstadt.de/iud/wwwmeth/lv/ss98/im2/monata2.htm dokumentiert.

3 Architektur und Entwurfsentscheidungen

Die Architektur des Informationsdienstes wird von folgenden Prinzipien determiniert:

3.1 Stabile „redundanzfreie" Strukur

WebSite 'Methodik weist eine recht konventionelle Struktur auf: Er ist in 5 Rubriken gegliedert, die weitgehend monohierarchisch weiterentwickelt werden. Allerdings gibt es in begründeten Fällen Dokumente, die auf zwei oder mehr Pfaden „top down" erreichbar sind, so daß ein näheres Hinsehen klar macht, daß es sich um eine Polyhierarchie handelt. Der logischen Gliederung in Rubriken entspricht weitgehend auch eine isomorphe Verzeichnisstruktur zur Ablage der Dateien.

Wichtigstes Prinzip bei der Ausarbeitung der Struktur ist es, daß die Einordnung eines Dokumentes in die Struktur zeitunabhängig und eindeutig sein muß. Triviales Beispiel: Ein aktueller Tagungsbericht kann nicht sinnvoll in einem Verzeichnis „Aktuelles" lokalisiert werden, denn eine solche Zuordnung ist zeitlich temporär. Eine Übung kann nicht Bestandteil einer Lehrveranstaltung sein, denn dieselbe Übung kann auch in anderen (oder folgenden) Veranstaltungen verwendet werden. Entsprechendes gilt für Beispiele, Skripten etc. Nach dieser Überlegung ist zwingend, daß die Inhaltsseite (konkrete Skripten, Übungen, Folien, Werkzeuge, Glossare) von ihrer Verwendung (in Lehrveranstaltungen oder etwa Vorträgen) zu trennen ist. In *WebSite 'Methodik* entspricht dies der Trennung in die Rubriken Lehrmaterialien (Abb. 7) und Aktivitäten. Die Lehrmaterialien umfas-

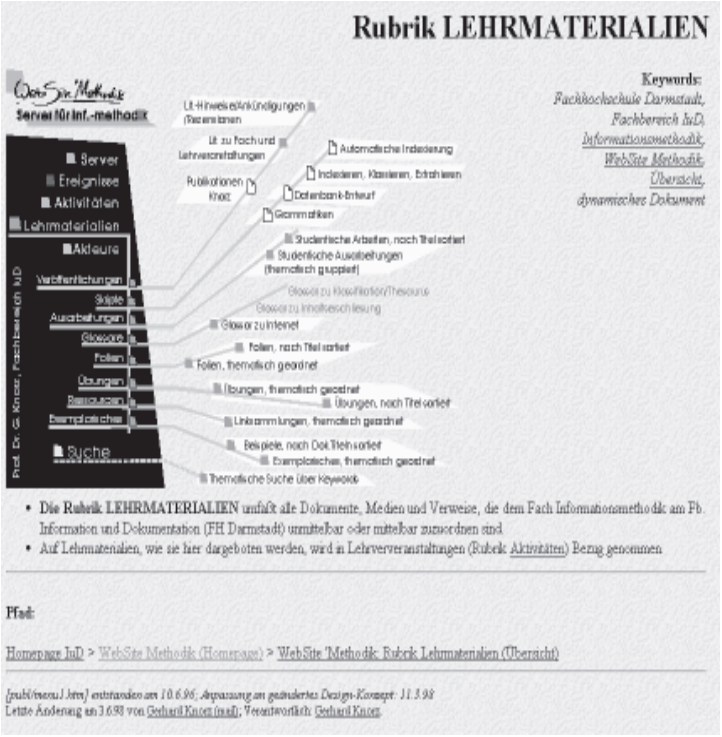


Abb. 7: *Übersichtskarte für die Rubrik Lehrmaterialien: 1 200 Dateien in 70 Verzeichnissen mit ca. 30 MB*

sen etwa 50 Prozent des insgesamt verfügbaren Informationsabgebotes. Die Struktur von *WebSite 'Methodik* hat sich insofern bewährt, als das rasche Wachstum des Servers an keiner Stelle eine konzeptionelle Reorganisierung notwendig gemacht hat.

3.2 „Monotonie-Eigenschaft" des Informationsangebotes

Ein wichtiges Prinzip von *WebSite 'Methodik* ist die Tatsache, daß im Detail ständig aktualisiert wird, daß laufend neue Dokumente hinzukommen, daß aber auf Dokumentenebene grundsätzlich nichts gelöscht wird: Ein gespeicherter Link auf ein Dokument bleibt valide, da dieses Dokument entweder zeitunabhängig von Interesse ist oder aber dem Ziel dient, einen in der Zeit ablaufenden Prozeß (z. B.

eine Lehrveranstaltung des Wintersemesters 96/97) zu dokumentieren.

3.3 Vollständigkeit

Ein zentrales Qualitätskriterium einer Datensammlung ist Vollständigkeit und Abdeckung. Eine Informationsressource, in der ein Kunde mehrfach Erwartetes nicht finden konnte, verliert entscheidend ihren Wert. Jeder kennt diesen Effekt von nicht gepflegten „Schwarzen Brettern", auf denen tatsächlich wichtige Ankündigungen nicht (mehr) wahrgenommen werden. In diesem Sinn gilt für *WebSite 'Methodik*: Wann immer eine Unterlage verwendet oder ein Aufsatz (wie etwa dieser hier) geschrieben wird, ist sie (bzw. er) in *WebSite 'Methodik* zu finden. Es ist also nie die Frage, *ob* etwas elektronisch verfügbar ist. Es ist evident, daß dieses Prinzip gerade in der Aufbauphase einen massiven Anforderungsdruck auf den Verantwortlichen ausgeübt hat.

3.4 Inhaltliche und formale Qualität

Eine studentische Gruppenarbeit mit schriftlichem Ergebnis ist zunächst unabhängig von ihrer Qualität als Teil der Lehrveranstaltungsdokumentation im Netz. Eine solche Arbeit gilt in *WebSite 'Methodik* allerdings nicht als (selbständiges) Dokument, sondern als Teil eines übergeordneten Dokumentes: Es taucht nicht in den einzelnen Registern auf und ist nicht indexiert. Selbstverständlich wird es bei Volltextsuchen gefunden. Erfüllt ein solches Dokument den von *WebSite 'Methodik* vertretenen Qualitätsstandard und ist es von allgemeinem Interesse, so wird eine Kopie dieses Dokuments in entsprechender „Verpackung" und mit einem Hinweis auf seine Entstehung in die Rubrikstruktur des Servers eingebaut. Diese Form der Anerkennung studentischer Arbeit bewirkt wahrnehmbar eine zusätzliche Motivation für Studierende.

Zur formalen Qualität gehört, daß jedes Dokument eine definierte Struktur mit obligatorischen Informationselementen besitzt: Indexierung, Titel, Angabe von Autor, Entstehungsdatum, Änderungsdatum, Verantwortlichkeit, Pfad- und Dateiname, Einbettung in die Dokumentenhierarchie und Standard-Navigationsmöglichkeiten. Dazu kommen die – nicht sichtbaren – Meta-Tags für Kurzfassung, Autor, Schlagwörter u. a. Der von den Browsern verwendete Titel (<TITLE>-Tag) folgt einheitlichen Regeln für die Benennung von Dokumenten, die dafür sorgen, daß in alphabetischen Listen zusammengehörende Dokumente zusammensortiert werden (siehe Abb. 7).

3.5 Navigation und Suche

Jedes Dokument ist durch seine Indexierung thematisch eingeordnet. Die verwendeten Deskriptoren führen per Link direkt in das Schlagwortregister, in dem thematisch verwandte Dokumente zusammengestellt sind. Alternativ werden (von jedem Dokument aus) Volltextsuche, ein alphabetisches Register und ein hierarchisches Register angeboten. Jedes Dokument listet darüber hinaus alle Standard-Pfade auf, über die dieses Dokument erreichbar ist. Ziel ist es, es dem Nutzer und seiner Zielsetzung zu überlassen, von welchen der angebotenen Navigationsmöglichkeiten er im einzelnen Gebrauch macht.

4 Weitere Entwicklung und Kontext

Die Internet-Technologien und -dienste sind in ständiger Entwicklung. Frames sind eine der (älteren) Möglichkeiten, von denen *WebSite 'Methodik* bisher keinen Gebrauch macht, XML ist eine aktuelle einschlägige Entwicklung, die wesentliche neue Möglichkeinen verspricht. Datenbankanbindung ist eine vordringliche Aufgabe, die strukturell bereits gut vorbereitet ist (jede Folie, jede Übung enthält eine kategorisierte Beschreibung im Stil eines Datenbankeintrags), die technisch kein Problem darstellt, die aber dennoch in der vollständigen Umsetzung einen beträchlichen Umstellungsaufwand bedeuten wird. Die in der Titelgrafik und in Abb. 7 gezeigten Übersichtskarten sind Teil einer gerade begonnenen Überarbeitung des Infodesigns von *WebSite 'Methodik*.

Inhaltlich erweitert sich die Perspektive, unter der *WebSite 'Methodik* betrieben wird, durch den neuen interdisziplinären Studiengang Media System Design an der FH Darmstadt, der den aktuellen Bereich der Entwicklung und Pflege multimedialer Systeme und Dienste adressiert. Der Fachbereich IuD wird in der Lehre die Bereiche Information Retrieval, Information Management vertreten und weitere Themen im zur Vertiefung und zur Schwerpunktsetzung einbringen.

Unter dem Aspekt neue Lehr- und Lernformen ist relevant, daß der Fachbereich gegenwärtig in mehreren beantragten Projektvorhaben vertreten ist, die Distance Learning insbesondere auch für Zwecke der Fort- und Weiterbildung zum Ziel haben. In dieser Hinsicht stellt *WebSite 'Methodik* eine Ressource dar, die einerseits als „Startkapital“ genutzt werden wird, die aber andererseits von den begleitenden Arbeiten und der entstehenden Infrastruktur auch profitieren wird.

CL-Demos im World Wide Web

Martin Volk, Universität Zürich
(e-mail: volk@ifi.unizh.ch)

Wie oft haben Sie schon nach einer Möglichkeit gesucht, mal eben schnell ein Morphologie-Programm oder einen Tagger für das Deutsche zu demonstrieren, wenn die Frage im Raum stand: Was ist denn eigentlich Computerlinguistik? Wie oft haben Sie schon bedauert, dass Sie ein Programm, von dem sie ausführlich gelesen hatten, nicht selbst ausprobieren können?

Durch die Anbindung von Programmen an das Internet können beide Probleme behoben werden. HTML-Formulare bieten die Möglichkeit, Wörter, Sätze oder ganze Texte in einem WWW-Browser einzugeben und über das Internet an ein verarbeitendes Programm weiterzuleiten. Viele Forscher machen auf diese Art mit online-Demo-Versionen auf ihre lauffähigen Programme aufmerksam, und Firmen bewerben so ihre Produkte.

Es ist jedoch bekanntlich schwierig, die gesuchte Stecknadel *‚NLP-Applikation*‘ im Heuhaufen ‚Internet‘ zu finden. Und bei dieser speziellen Aufgabe helfen auch die Internet-Suchdienste nicht viel, denn interaktive NLP-Seiten finden sich unter sehr unterschiedlichen Bezeichnungen (wie z.B. „HPSG-Interaktiv“ = ein Parser; „Interactive Demo of SVOX“ = ein Sprachsynthesesystem; „NLP Parser Demonstration Page“ = eine Auswahl unter mehreren Parsern und Parsingstrategien). Deshalb haben wir eine Sammlung mit **Interactive online CL demos** angelegt unter der Adresse: http://www.ifi.unizh.ch/CL/InteractiveTools.html.

Alle dort aufgeführten WWW-Links verweisen auf interaktive sprachverarbeitende Programme im Internet. Dabei bedeutet ‚interaktiv‘, dass der Benutzer eine Eingabe im WWW-Browser machen kann und diese durch das verarbeitende Programm unmittelbar analysiert und das Ergebnis zurückgeschickt wird. Unsere Sammlung beschränkt sich auf Systeme für Deutsch und Englisch und umfasst:

- Morphologie-Analyse von Wortformen
- Wortarten-Tagger zur kontext-abhängigen Zuteilung von Wortarten für einen oder mehrere Sätze
- Parser zur Syntaxanalyse einzelner Sätze
- Transformations-Systeme zur Umwandlung eines Aktiv-Aussage-Satzes in einen Passiv-Satz oder in einen Frage-Satz

- Generierung von Sätzen und kleinen Texten aus z. B. Börsentabellen oder Bildbeschreibungen
- Synthese gesprochener Sprache aus geschriebener Sprache
- Frage-Antwort-Systeme z. B. über geographische Fakten
- Maschinelle Übersetzung von einigen Sätzen oder ganzen WWW-Seiten

Programme, die lediglich das Nachschlagen in einem Lexikon ermöglichen, wurden nicht aufgenommen, jedoch enthält die Sammlung einige Systeme für online-Korpusrecherchen (unter Auflösung von Flexion und anderen Wortbildungen) und für Recherchen in semantischen Netzwerken. Verweise auf Grammatik- oder Stilprüfprogramme fehlen, da dazu bisher keine interaktiven WWW-Angebote gefunden werden konnten.

Schließlich gibt es noch einen Abschnitt mit „General Tools“. Dort finden sich z. B. Sprachidentifizierungsprogramme, also Programme, die aufgrund eines oder mehrerer eingegebener Sätze entscheiden können, um welche Sprache es sich handelt.

Wir hoffen, dass die Aufstellung für viele von Ihnen nützlich ist. Um die Sammlung aktuell halten zu können, sind wir darauf angewiesen, dass Sie uns Hinweise auf neue interaktive CL-Software (und natürlich auch auf alte ungültige Links) schicken. Bitte senden Sie Ihre Vorschläge an Martin Volk, Universität Zürich (volk@ifi.unizh.ch).

Nachrichten

Habilitation im Fach Computerlinguistik

Dr. Nico Weber, Mitglied der GLDV und Leiter des Arbeitskreises „Lexikographie", hat am 13. Februar 1998 seine Habilitation vor der Philosophischen Fakultät der Universität Bonn abgeschlossen und die Venia Legendi für das Fach Computerlinguistik erworben. Die Habilitationsschrift hat das Thema *„Die Semantik von Bedeutungsexplikationen"* und wird in Kürze im Verlag Peter Lang erscheinen.

Die GLDV gratuliert herzlich!

Projekte

Quantitative Verfahren zur Zuordnung von Präpositionalphrasen

Hagen Langer (Universität Osnabrück), Stephan Mehl (Universität Duisburg), Martin Volk (Universität Zürich)

Im September 1996 fand im Rahmen der KONVENS in Bielefeld unter der Leitung von Andreas Mertens und Marion Schulz (FernUniv. Hagen) sowie von Stephan Mehl ein Workshop zum Problem der PP-Zuordnung statt. Im Anschluß an diesen Workshop bildeten Stephan Mehl, Hagen Langer und Martin Volk eine Arbeitsgruppe mit dem Ziel, die Effizienz quantitativer Verfahren der PP-Zuordnung für das Deutsche zu untersuchen. Angeregt durch die Ergebnisse von Hindle/ Rooth 1993 sind in den letzten Jahren zahlreiche Arbeiten entstanden, die das Problem der PP-Zuordnung für das Englische mit statistischen Mitteln erfolgreich angehen. Diese Ergebnisse sind aber aufgrund der freieren Wortstellung nicht ohne weiteres auf das Deutsche übertragbar. Außerdem setzen viele einschlägige Arbeiten eine umfangreiche Treebank voraus, die für das Deutsche noch nicht existiert.

Ausgangspunkt der gemeinsamen Untersuchungen ist die Tatsache, daß Valenzerwartungen eine bedeutende Rolle bei der syntaktischen Disambiguierung spielen. In vielen Fällen kann eine Entscheidung über die korrekte PP-Zuordnung dadurch getroffen werden, daß entweder das Verb oder eines der möglichen Bezugsnomina eine Präpositionalphrase mit der betreffenden Präposition erwartet. Unter „Erwartung" sind dabei in erster Linie obligatorische oder fakultative Valenzstellen zu verstehen, darüber hinaus aber auch statistisch signifikante Kookkurrenzen wie der Ausdruck „Familien mit Kindern" u. ä. Leider ist für das Deutsche kein maschinenlesbares Lexikon mit vollständigen Valenzangaben verfügbar. Freundlicherweise stellte das Zentrum für elektronische Ressourcen europäischer Sprachen (ZERES) in Bochum uns ein Lexikon zur Verfügung, das für einen Teil der deutschen Verben Valenzangaben auflistet. Die in Nijmegen entwickelte CELEX-Datenbank enthält ebenfalls zahlreiche Valenzangaben. Insbesondere für Nomina fehlen uns solche Angaben jedoch. Darüber hinaus gibt es Sätze, in denen sowohl Verb als auch Nomina konkurrierende Erwartungen an eine PP stellen. Für solche Fälle wäre eine numerische Gewichtung der Valenzangaben interessant.

Unsere bisherigen Resultate bestätigen, daß quantitativ ermittelte Erwartungswerte auch für das Deutsche einen wichtigen Beitrag zur Disambiguierung leisten können. Allerdings ist es für weniger häufige Lexeme und für solche mit fakultativen Valenzstellen schwierig, genügend Belege zu finden, um Valenzstellen automatisch zu bestimmen bzw. quantitativ zu gewichten. Weitere Schwerpunkte unserer Arbeit neben der Valenzproblematik sind u. a. die Ermittlung fester Wendungen und die Einbeziehung grober semantischer Informationen, z. B. die Identifizierung von Zeitangaben. Hagen Langer untersucht darüber hinaus den Einsatz von Bindungswahrscheinlichkeiten in einem probabilistischen Parser. Ein weiterer Aspekt ist die Untersuchung fachtextspezifischer Unterschiede.

Zwischenbericht: Langer, Hagen/Mehl, Stephan/Volk, Martin: „Hybride NLP-Systeme und das Problem der PP-Anbindung." In: Workshop „Hybride konnektionistische, statistische und symbolische Ansätze zur Verarbeitung natürlicher Sprache", KI 97, Freiburg, http://www.dfki.de/~busemann/ki97/ki97-ws03.html.

Literatur: D. Hindle/M. Rooth (1993): Structural ambiguity and lexical relations. Computational Linguistics 19(1), 103–120.

Evaluation maschineller Übersetzungssysteme

Ein Projekt des Arbeitskreises „Maschinelle Übersetzung“ der GLDV

Uta Seewald

Seit Beginn des Jahres 1998 liegt der Schwerpunkt des Arbeitskreises „Maschinelle Übersetzung“ auf der Evaluierung maschineller Übersetzungssysteme. Das Evaluationsverfahren ist auf zwei Veranstaltungen des Arbeitskreises in Saarbrücken, am 30. Januar 1998 und am 8. Mai 1998, ausgearbeitet worden.

Auf der Veranstaltung am 30. Januar ging es zunächst um die Erarbeitung eines Kriterienkatalogs zur Durchführung der Evaluation. Um hierbei möglichst vielfältige Aspekte und Erfahrungen mit Evaluationsvorhaben zu berücksichtigen, berichteten die Teilnehmerinnen und Teilnehmer im ersten Teil der Veranstaltung über bereits durchgeführte Evaluationen oder einzelne Aspekte solcher Evaluationen. Der zweite Teil des Arbeitskreistreffens war schließlich der eigentlichen Erarbeitung des Kriterienkatalogs zur Evaluation von Übersetzungssystemen gewidmet. Der hier erarbeitete Kriterienkatalog sieht ein zweistufiges Evaluationsverfahren vor: die erste Stufe der Evaluation wird ausschließlich kommerziell vertriebene Übersetzungssysteme berücksichtigen. In der zweiten Stufe sollen dann in der Forschung im Einsatz bzw. in Entwicklung befindliche Systeme evaluiert werden, wobei insbesondere der Frage nachgegangen werden soll, was Forschungssysteme in ihrer Leistung gegenüber kommerziell vertriebenen Systemen auszeichnet. Um die in ihrem Aufbau, Umfang und hinsichtlich der Benutzerfreundlichkeit so unterschiedlichen Systeme, wie sie im Fall der kommerziellen und der Forschungssysteme vorliegen, vergleichen zu können, erschien es sinnvoll, das Schwergewicht der Evaluation auf linguistische Kriterien zu legen, da diese gleichermaßen an kommerziellen und an Forschungssystemen überprüft werden können.

Um sicherzustellen, daß die Evaluationsergebnisse in bezug auf die linguistische Abdeckung der Systeme tatsächlich vergleichbar sind, wurde ferner festgelegt, in der ersten Evaluationsphase ausschließlich Systeme mit dem Sprachpaar Englisch – Deutsch und zwar mit Englisch als Quell- und Deutsch als Zielsprache zu berücksichtigen: *T1 Professional* (Langenscheidt, GMS), *Personal Translator Plus ’98* (Linguatec/Rheinbaben & Busch, IBM), *Power Translator Pro* (Globalink), *SYSTRAN PROfessional für Windows* (Systran), *Transcend* (HEI-Soft,

Intergraph), *Logos* (Logos). Systeme der untersten Preisklasse wurden aufgrund ihrer in verschiedenen bereits durchgeführten Testläufen als unbefriedigend eingestuften Übersetzungsqualität von vornherein nicht einbezogen.

Aus einer Reihe von linguistisch für das betreffende Sprachpaar als relevant eingestuften Phänomenen wurden zunächst imperativische Strukturen, idiomatische Wendungen sowie die Erkennung von Komposita zu einer Recherche und Analyse an den für das Testkorpus vorgesehenen Textsorten ausgewählt. Ausschlaggebend für die Zusammenstellung des Testkorpus war die Frage, welche Textsorten aufgrund ihrer sprachlichen Struktur für eine maschinelle Übersetzung geeignet sind bzw. im professionellen Umfeld häufig mit maschineller Unterstützung übersetzt werden. Von den zunächst ausgewählten Textsorten bzw. Textsortenklassen, die sich aus Handbüchern, Instruktionstexten, Firmenjahresberichten, Handelskorrespondenz, Web-Seiten von Reiseanbietern sowie Web-Seiten aus dem Bereich des *Electronic Commerce* zusammensetzten, wurden auf dem Arbeitstreffen am 8. Mai auf der Basis der in der Zwischenzeit durchgeführten Textrecherchen und linguistischen Untersuchungen schließlich Instruktionstexte, d. h. Reparaturanweisungen aus der Automobilbranche und Softwareinstallationsanleitungen, sowie Web-Seiten aus den Bereichen Tourismus und *Electronic Commerce* ausgewählt.

Um die Vergleichbarkeit der in die Evaluation einbezogenen Systeme sowie eine möglichst hohe Transparenz in bezug auf die Testergebnisse zu gewährleisten, beschränkt sich die Evaluation zum einen ausschließlich auf linguistische Phänomene. Zum anderen ist die angestrebte Vergleichbarkeit und Transparenz auch der Grund, warum im Rahmen der ersten Evaluationsphase *Translation Memories* bzw. Satzarchive nicht einbezogen werden, selbst wenn diese bereits als Module einzelner Systeme angelegt sind. – Die Festlegung auf ein solches Verfahren führte schließlich dazu, idiomatische Wendungen aus dem Katalog der linguistischen Phänomene auszuklammern und stattdessen Konditionalsätze und syntaktische Koordinationen in die Untersuchung einzubeziehen. Die Evaluation der einzelnen linguistischen Phänomene wird jeweils von einem Evaluator bzw. einer Evaluatorin hauptverantwortlich durchgeführt, wobei alle Phänomene anhand von jeweils 300 Testsätzen an jedem der ausgewählten maschinellen Übersetzungssysteme überprüft werden.

Die Bewertung der grammatischen Korrektheit der maschinellen Übersetzung der ausgewählten linguistischen Phänomene soll anhand eines viergliedrigen Klassifikationsschemas erfolgen, das die Bewertungspunkte „Satz bzw. Syntagma vollständig korrekt“, „Satz bzw. Syntagma in bezug auf das zu überprüfende

linguistische Phänomen grammatisch korrekt“, „Satz bzw. Syntagma hinsichtlich des zu überprüfenden linguistischen Phänomens falsch“ und „ Satz bzw. Syntagma falsch; Fehlerursache nicht eindeutig entscheidbar“ umfaßt.

Die Ergebnisse der ersten Evaluationsphase, die auf den Arbeitskreistreffen im Januar und im Mai initiiert wurde, sollen im Rahmen eines Workshops auf der Konvens ’98 einem größeren Publikum vorgestellt und mit Anwendern maschineller Übersetzungssysteme, Vertretern aus dem Bereich der Forschung und dem industriellen Entwicklungsumfeld diskutiert werden.

Wolfgang Dalitz und Gernot Heyer:

HYPER-G. Das Informationssystem der 2. Generation.

dpunkt Verlag für digitale Technologie GmbH, Heidelberg 1995
DM 68,-

Besprochen von Rita Nübel, Saarbrücken (e-mail: rita@iai.uni-sb.de).

Die beiden Autoren stellen in ihrer Dokumentation die wichtigsten Eigenschaften und Funktionalitäten des HYPER-G-Informationssystems vor, das sie selbst in Abgrenzung zu Informationssystemen der ersten Generation (wie WWW, Gopher und WAIS) als System der zweiten Generation einordnen. HYPER-G soll die Mängel und Schwächen der Vorgänger vermeiden (S. 19), wie z. B. deren mittlerweile unzureichende Suchmechanismen, die am rasant zunehmenden Informationsangebot im Internet mangels Effizienz scheitern. Gleichzeitig ist den Autoren allerdings auch klar, daß dieses Buch im Prinzip nur eine „[...] Momentaufnahme des jetzigen Zustands, vom August 1995 [...]“ (S.1) sein kann, die die möglichen Veränderungen und Weiterentwicklungen, die unumstrittenerweise zu erwarten sind, nur teilweise antizipieren kann. Somit sind zukünftige HYPER-G-Anwender zumindest schon mal vorgewarnt: Funktionalitäten, die in den einzelnen Kapiteln meist sehr präzise und grafisch ansprechend dargestellt sind, können im realen System etwas anders aussehen, bzw. ganz fehlen.

Das Buch ist in zehn Kapitel gegliedert, in denen die Autoren versuchen, einen breiten Bogen von der Standardeinführung über Informationssysteme im Internet bis hin zur Beschreibung der UNIX-Schnittstelle, Verwaltung und Installation des HYPER-G Servers bzw. Clients zu spannen. Hierbei sind Adressaten mit unterschiedlichem technischen Vorwissen angesprochen (normaler Anwender, Systemadministrator).

Nach einer sehr kurzen Einleitung (Kap. 1) faßt das zweite Kapitel die grundlegenden Konzepte des Internet sowie seine Informationssysteme (z. B. Gopher, FTP, WWW) und Software-Werkzeuge für den Datenaustausch zusammen. Große Nachteile bei diesen Systemen, so die Autoren, liegen in der Konzeption der Strukturierung des Informationsangebots, den bekannten Inkonsistenzen bei der Verwendung von URLs sowie der noch nicht oder nur rudimentär vorhandenen Unterstützung des immer mehr aktiv agierenden Lesers („[...] jeder Leser (ist) gleichzeitig auch Autor [...]“, S. 19). Vor- und Nachteile von Suchwerkzeugen

(Archie, Veronica, WAIS) für gezielte Suche nach Daten werden ebenfalls beschrieben.

Das dritte Kapitel beginnt mit einer kurzen Darstellung der Hauptprobleme von Hypertext-Systemen der ersten Generation. Der Einstieg in HYPER-G erfolgt dann über die Beschreibung seiner wichtigsten Eigenschaften. Im einzelnen sind dies Orientierungs- und Navigationshilfen, Volltextindexierung, bidirektionale Links als eigenständige Objekte (Objektorientiertheit),Verbindungsmöglichkeiten zwischen verschiedenen Typen von Dokumenten (Ton, Bild, Video, 3D-Objekte), Interaktivität, Zugriffskonzept für Benutzer und Gruppen, Cache-Verwendung, Kompatibilität mit anderen Internet-Informationssystemen (Interoperabilität) und Mehrsprachigkeit. Daß HYPER-G neben den fünf europäischen Hauptsprachen (Englisch, Französisch, Deutsch, Italienisch und Spanisch) auch Steyrisch unterstützt, erfährt der interessierte Leser allerdings erst im fünften Kapitel. Das Designkonzept von HYPER-G wird generell eingeführt und im einzelnen am Beispiel des HYPER-G Clients Harmony erläutert. Relativ spät werden die bereits vorher eingeführten Begriffe „Objekt“ und „Attribut“ im Bezug auf die Objektorientiertheit von HYPER-G definiert.

Das vierte Kapitel erläutert die Grundlagen der Bedienung von HYPER-G am Beispiel der Clients Harmony (für X11/Unix) und Amadeus (für PC/Windows). Die Bedienung der Clients ist ähnlich, nur ist Amadeus weniger aufwendig realisiert. Verschiedene Funktionalitäten wie beispielsweise das Navigieren in Dokumentmengen oder entlang der Hyperlinks, Datenbanksuche, Suchbegriffe, Sprachauswahl (nur in Harmony) oder die Suche in Textdokumenten, Darstellung und Ausgabe von Dokumenten sowie Benutzereinstellungen werden ausführlich für jeden der Clients beschrieben. Die Suche nach bestimmten Begriffen in Titeln, Stichworten und im Inhalt von Hypertext- „[...] und bald auch in Postscript-Dokumenten“ (S.54) wird ebenfalls dokumentiert.

Im umfangreichen fünften Kapitel beschreiben die Autoren den Aufbau eines Hypermedia-Informationsangebots mit HYPER-G. Nachdem sie ein paar generelle Tips zur Erstellung eines Informationangebots gegeben haben, erläutern sie sehr detailliert die wichtigsten Grundbegriffe und Funktionen jeweils wieder für die Clients Amadeus und Harmony.

Unter anderem wird das Konzept der Kollektionshierarchie für die Einordnung von Dokumenten in Dokumentmengen in HYPER-G definiert. Danach schließt sich eine Beschreibung der Hypertextformate HTML und (speziell für HYPER-G) HTF an, wobei die Grundstruktur von HTF-Dokumenten erläutert wird. Die Erzeugung und Bearbeitung von Dokumenten, das Einfügen von Hypertext- und Mul-

timedia-Dokumenten, das Erstellen von Verweisen auf Fremddokumente (Telnet-Verbindungen, Gopher-Menüs und Dokumente, WWW-Dokumente), die Funktion von Hyperlinks und Attributen sowie Anleitungen für den Aufbau eines mehrsprachigen Informationssystems in HYPER-G wird ausführlich dokumentiert. Nützlich sind hier die Auflistung möglicher Fehlermeldungen vom System und jeweils Hinweise, mit denen man das Problem löst.

Das sechste Kapitel beschreibt UNIX-Kommandos für die Manipulation des HYPER-G Servers über selbst zu schreibende Scripts. Im siebten Kapitel werden Benutzer- und Gruppenverwaltung in HYPER-G dokumentiert, Kapitel acht erläutert die einzelnen Installationsschritte für den HYPER-G Server und das neunte Kapitel illustriert die Installation der HYPER-G Clients am Beispiel von Amadeus. Im Schlußkapitel findet der Leser Informationen zu HYPER-G- und Perl-Quellen sowie eine Kurzdokumentation der beigelegten CD-ROM.

In diesem Zusammenhang noch ein letztes Wort zum *information access* im ganz konkreten Sinn: Bei dem Versuch, die dem Buch beigelegte CD-ROM auszupacken, zerreißt man unweigerlich die Innenseite des Buchdeckels, was den Buchliebhaber natürlich verärgert. Vielleicht könnte man die nächste Ausgabe als online-Dokumentation direkt mit auf die CD-ROM brennen, damit könnten dann auch gleichzeitig die neuesten Entwicklungen seit Erscheinen des Buches aufgearbeitet werden.

Insgesamt ist den Autoren eine recht übersichtliche und gut strukturierte Dokumentation mit Handbuch-Charakter gelungen, die zumindest im Jahre 1995 ein breites Publikum von HYPER-G-Anwendern bedienen konnte.

Guillaume Schiltz:

Der Dialektometrische Atlas von Südwest-Baden (DASB). Konzepte eines dialektometrischen Informationssystems.

= *Studien zur Dialektologie in Südwestdeutschland*

5. Auflage. Hrsg. v. Hugo Steger, Eugen Gabriel und Volker Schupp
Marburg: N. G. Elwert Verlag, 1996

Besprochen von Reinhard Köhler, Trier.

Die Dialektometrie ist eine exakte, auf einer Reihe von Explorationsverfahren sowie quantitativen mathematischen Methoden wie numerischer Taxonomie und der Analyse von Wahrscheinlichkeitsverteilungen beruhende dialektologische Teildisziplin, die vor allem dank der bekannten Arbeiten von Hans Goebl (Salzburg) methodologisch außergewöhnlich gut reflektiert, wissenschaftstheoretisch fundiert und praktisch bewährt ist. Dieses induktive Verfahren dient der Aufbereitung und Analyse der in Sprachatlanten verfügbaren Daten und umfaßt die Schritte *Erhebung*, *Merkmalsbestimmung*, *Taxierung*, *Ähnlichkeitsmessung*, *Auswertung* und *Visualisierung* (Kartographie). Das Forschungsgebiet ist in mehrfacher Hinsicht wissenschaftlich (vor allem methodologisch und methodisch) sehr anspruchsvoll und technisch aufwendig, letzteres einerseits wegen der rechenintensiven statistischen Methoden und der großen Menge der zu handhabenden Daten[1] und andererseits aufgrund der hohen Bedeutung der visuellen Aufbereitung der Analyseergebnisse in Form von verschiedenen Kartentypen. Dieser technische Aufwand macht den Einsatz von Rechenanlagen und spezieller Software unabdingbar.

Entwickelt wurde die Dialektometrie innerhalb der romanistischen Dialektologie, deren Sprachatlanten traditionell die Rohdaten wiedergeben, so daß der dialektometrischen Analyse sämtliche denkbaren Merkmale zur Auswahl stehen, die in den Sprachbelegen enthalten sind. Solche Sprachatlanten sind in der Regel Wort-bezogen und geben die Belege in phonetischer Transkription wieder, so daß je nach Untersuchungsziel verschiedene lexikalische, morphologische, phonologische und phonetische Merkmale betrachtet werden können. Demgegenüber sind in Atlanten der germanistischen Dialektologie bereits Vorauswahlen getroffen. Nicht die Originaldaten werden veröffentlicht – vielmehr enthalten diese Atlanten symbolische Darstellungen der Ausprägungen von den als relevant

angesehenen Merkmalen der betreffenden Wörter.

Guillaume Schiltz hat mit seinem *Dialektometrischen Atlas von Südwest-Baden* in Form von einem Textband (Beilage: Grundkarten-Folie mit den Meßpunktnummern, drei Clusterprotokolle sowie eine Diskette, s. u.) und drei Kartenbänden den erfolgreichen Versuch vorgelegt, die dialektometrischen Verfahren unter Berücksichtigung der germanistischen Arbeitsweise auf einen Teil des *Südwestdeutschen Sprachatlas* anzuwenden.

Der Textband beginnt mit einer kurzen einleitenden Darstellung von Zielsetzung und Methodik und erläutert zunächst die Grundideen und Prinzipien der numerischen Taxonomie, die vor allem in der Biologie entwickelt worden ist (und übrigens von G. Altmann und W. Lehfeldt 1973 in die *Sprachtypologie* eingeführt wurde). Die dialektometrischen Grundlagen werden auf 50 Seiten eingehend und detailliert dargestellt, wobei Schiltz sich an die Abfolge der Arbeitsschritte (s. o.) hält. In mustergültiger Weise werden alle wichtigen Probleme, Entscheidungen, Prinzipien, Verfahren, Maße und Algorithmen explizit vorgestellt und erörtert – wie es H. Goebl zu Recht vorführt und verlangt; man erfährt sogar, welche Techniken bei der Visualisierung und der kartographischen Gestaltung verwendet wurden (und warum).

Auf 30 Seiten werden Ziel, Datenmaterial und Taxation des *Dialektometrischen Atlas* zusammengefaßt – betrachtet werden ausschließlich phonetische Merkmale (Vokalqualitäten und -quantitäten) – und die resultierenden Karten kommentiert. Die in den drei Kartenbänden wiedergegebenen Kartentypen umfassen Zwischenpunktkarten, Ähnlichkeitskarten, Clusterkarten und Kennwertsynopsen (Maxima, Kommunikationsgüten und Interaktionsprodukte). Abschließend kommt Schiltz in einer dialektometrisch-typologischen Interpretation zu einem Vorschlag[2] zur Gliederung des Dialektraums (für die es bislang keine allgemein akzeptierte Lösung gibt).

Zu allen Kartentypen gibt der Autor Erläuterungen und bespricht mehrere Beispiele im einzelnen. Mit den Karten selbst kommt man sofort zurecht, da Schiltz den von Goebls Arbeiten bekannten Konventionen folgt.

In seinem Resümee kommt der Autor zu dem Schluß, daß die etablierten dialektometrischen Verfahren mit geringen Anpassungen an die andersartige Datenbehandlung in germanistischen Atlanten erfolgreich auch in der germanistischen Dialektologie verwendet werden können und zu objektiveren Resultaten führen als es mit den diachronisch-vergleichenden oder strukturalistischen Methoden möglich ist – ein Fazit, dem man nur zustimmen kann.

Besonders erfreulich ist es, daß Schiltz sich entschlossen hat, der Publikation seiner Arbeit eine Fassung der selbsterstellten Software beizulegen: eine Diskette mit MS-DOS™-Programmen und einer Auswahl seiner Daten, kurz: ein dialektometrisches Geo-Informationssystem mit der Bezeichnung *DIALECTOm*. Dieses System, ergänzt durch ein *TUTORAT*, lädt zum praktischen Nachvollziehen der Messungen, Berechnungen und Visualisierungen in verschiedensten Variationen von Parametern bzw. Verfahren geradezu ein und erlaubt es auch, mit anderen Daten neue dialektometrische Studien durchzuführen. Die Handhabung ist völlig problemlos; außerdem sind Aufbau und Arbeitsweise im Textband gut dokumentiert.

Insgesamt ist Schiltz' Atlas weit mehr, als der Titel verspricht: eine Einführung in und ein Exempel für die dialektometrische Arbeitsweise, der breite und gründliche Rezeption zu wünschen ist. Vielleicht verführen die klare Argumentation, die sichtbaren Resultate und – nicht zuletzt – die fertige, mitgelieferte Software auch den einen oder anderen germanistischen Dialektologen dazu, mit den Methoden der Dialektometrie zu experimentieren.

Anmerkungen

[1]Ein Sprachatlas umfaßt typischerweise mehrere hundert Seiten, jede Seite kann mehrere hundert Meßpunkte enthalten, und aus den Angaben einer Seite zu den Meßpunkten können wiederum mehrere Merkmale abgeleitet werden.

[2]Als Nicht-Dialektologe ist der Rezensent nicht in der Lage, dieses Ergebnis zu beurteilen.

Literatur

Altmann, G. & Lehfeldt, W.: Allgemeine Sprachtypologie. München: Fink, 1973.

Goebl, H.: Eléments d'analyse dialectométrique (avec application à l'AIS). In: Revue de linguistique romane 45, 349 – 420.

Goebl, H.: Dialectometry: A short overview of the principles and practice of quantitative classification of linguistic atlas data. In: R. Köhler & B. Rieger [Hrsg.]: Contributions to Quantitative Linguistics. Dordrecht, Boston, London: Kluwer, 1993, 277 – 315.

Schiltz, G.: Dialektometrische Untersuchungen des schwäbisch-alemannischen Übergangsgebietes. In: Akten des IX. Internationalen Germanisten-Kongresses Vancouver 1995, Band 3 (Abstracts). Tübingen: Niemeyer, 1996, 62.

Schiltz, G.: Der Dialektometrische Atlas von Südwest-Baden (DASB). In: A. Ruoff & P. Löffelad [Hrsg.]: Syntax und Stilistik der Alltagssprache. Beiträge der 12. Arbeitstagung zur alemannischen Dialektologie. Tübingen: Niemeyer, 1997, 231 – 233.

Feldweg, H. und Erhard W. Hinrichs (Hrsg.):

Lexikon und Text. Wiederverwendbare Methoden und Ressourcen zur linguistischen Erschliessung des Deutschen. *[SI]*

Lexicographica Series Maior 73
Tübingen 1996, Niemeyer Verlag

Hoetker, W. und Petra Ludewig (Hrsg.):

Lexikonimport, Lexikonexport. Studien zur Wiederverwendbarkeit lexikalischer Informationen. *[LSM]*

Sprache und Information, Bd. 31
Tübingen 1996, Niemeyer Verlag

Rezensiert von Johann Haller, Saarbrücken (e-mail: hans@iai.uni-sb.de).

Gleich zwei Bände mit fast gleichem (Unter-)titel zur gleichen Zeit im gleichen Verlag: Das muß doch ein wichtiges Thema sein – oder hat da einer bei Niemeyer nicht aufgepaßt? Aber bei näherem Hinsehen ergibt sich, daß nur ein Artikel gleichzeitig in beiden Bänden zu finden ist, und immerhin einmal in Englisch und einmal in Deutsch.

Aus welchen theoretischen und praktischen Gründen das Thema der Wiederverwendbarkeit so wichtig ist, erfährt man in sehr übersichtlicher und systematischer Weise in der Einleitung von SI, das neun längere und in drei nach Forschergruppen getrennte Aufsätze enthält. Zu einem theoretischen Umschwung (mehr lexikalistisch begründete Sprachtheorien) kommt die Zunahme der praktischen Relevanz lexikalischer Ressourcen, die sich auch in der steigenden Anzahl von EU-Projekten zu diesem Thema zeigt. Einige davon (EAGLES, DELIS, MULTEXT, ET10-51) sind zwar dem Spezialisten bekannt, es ist aber sehr verdienstvoll, daß gerade die Ergebnisse dieser Projekte ausführlich und im Zusammenhang in deutscher Sprache zugänglich gemacht werden. Gerade die mangelnde Sichtbarkeit dieser und ähnlicher Projekte in der Öffentlichkeit war mehrmals kritisiert worden. Das aktuelle mehrsprachige Korpus- und Lexikonprojekt der EU (PAROLE) sollte eigentlich die Fortsetzung sein; daß die Fortwirkungen dieser Projekte sich besonders im deutschen Teilprojekt (PAROLE-D) in Grenzen halten, mag zunächst

darauf zurückzuführen sein, daß sich die genannten EU-Projekte hauptsächlich mit der englischen Sprache beschäftigten. Hier und da taucht mal ein deutsches Beispiel zwischen den „be"-Formen und den syntaktischen Verwendungsmöglichkeiten von „taste" auf. Außerdem sind zwei Jahre in der Computerlinguistik ein langer Zeitraum (die entsprechenden Workshops fanden 1994 statt), in dem Ergebnisse schnell veralten. Beide Bände (SI und LSM) machen eigentlich neugierig auf die weiteren Fortschritte; das studentische Beispielprojekt in einem Osnabrücker SI-Artikel, das sich mit der Anwendung des mehrfach verwendbaren Lexikons in Sprachlernsystemen befaßt, hofft darauf, daß die (eigentlich altbekannten) Schwierigkeiten bald überwunden werden: Disjunktions- und Negations-Operatoren für den Formalismus, Anbindung an eine große lexikalische Datenbank für partielle Lexikoneinträge etc. Ein wichtiger Schritt in diese Richtung ist sicher mit der in den drei Bochumer Beiträgen beschriebenen Bearbeitung des Cobuild-Lexikons geschehen; dieser wird von Schnelle auf der Hypothese der Bedeutung als Implikation theoretisch fundiert, und von den beiden anderen Autoren am Beispiel des HPSG-Formalismus ausgearbeitet und technisch umgesetzt. Ein Teil des Ergebnisses ist (noch kostenfrei) im WWW zu besichtigen (http://ww.linguistics.ruhr-uni-bochum.de/ccsd/).

Einen diffuseren Eindruck macht zunächst LSM, das eine Reihe kürzerer Artikel enthält; die in der kurzen Einleitung unternommene Gliederung in sechs Bereiche hilft da nur wenig weiter: Lexikon-Text, Wörterbücher, Wissensrepräsentation, Korpora, Linguistische Annotation, Statistische Disambiguierung. Zur lakonischen Kürze kommt auch noch ein Druck- bzw. Grammatikfehler in der ersten Zeile. Um die formale Kritik gleich auf einmal loszuwerden: Druckfehler, leider auch sinnentstellende, gibt es eine ganze Menge. Daß mit „Beispielgruppenabgabe" eigentlich die entsprechende „Angabe" gemeint ist, bedarf schon einiger Überlegung, bei „müßen" mit scharfem s war wohl die neue Rechtschreibung irgendwie schuld. Und schließlich meinen noch ein paar Autoren, sie müßten ihre Lateinkenntnisse an den Mann bringen. Auch hier kann man „computativ" noch erschließen, aber bei „Fehlende Lexika sind ein Desiderat" hätte man es auch einfacher haben können. Vom Inhalt her findet man in manchen Artikeln Interessantes: immer noch die Bonner Wortdatenbank (eine Pionierleistung, wenn auch inzwischen ein bißchen betagt), endlich ein paar Lichtblicke aus dem Institut für deutsche Sprache in Mannheim (jetzt mit dem Projekt COSMAS auch im Internet zu erproben: http://www.ids-mannheim.de/ldv/cosmas/), von dem man schon seit vielen Jahren computerlinguistische Aktivitäten erhofft hätte, Information über das eindrucksvolle CISLEX in München (die deutsche Entsprechung zum

LADL-Lexikon von Maurice Gross in Frankreich), Information zu einem großen Korpus-Projekt der EU (MULTEXT) und für Programmierfans eine Reihe technischer Beschreibungen, wie man zu einsatzfähigen Programmen kommt, die auch mit großen Text- und Wortmengen umgehen können. Schwedisch-Fans werden auch bedient; für diese Sprache – und dann auch für Deutsch – werden Vorschläge für Tagsets gemacht, über die aber die Zeit vermutlich schon wieder hinweggegangen ist. Daß doch gelegentlich auch auf alte Materialien zurückgegriffen wird, zeigen die vereinzelten Verweise auf alte Saarbrücker Wörterbücher, Analyse- und Lemmatisierungsprogramme, die bei der Konfektion umfangreicher Vorlagen zur Anwendung kommen. Solche werden bei den Versuchen des stochastischen Tagging benötigt, deren Beschreibung den Abschluß von LSM bildet. LSM hat am Schluß einen lieblos gemachten (und damit ein bißchen nutzlosen) Index, in dem ‚Wiederverwertbarkeit' gerade einmal auftaucht, dafür „leere" Wörter wie ‚Zeichen', ‚ZEIT' und einige Varianten linguistischer Begriffe wie ‚Wortartklassifikation', ‚Wortartenklassifikation' etc. Hier wäre es sicher angebracht, sich einmal auf eine Schreibweise festzulegen und diese dann beizubehalten. Da hätte man es lieber wie SI machen sollen und den Index ganz weglassen.

Insgesamt betrachtet, kann der fortgeschrittene Student der Computerlinguistik oder der Forscher, der etwas über den Stand der maschinellen Lexika in Deutschland wissen will, durch die Lektüre der beiden Bände einen recht guten und vollständigen Überblick bekommen; es fragt sich natürlich, ob in Zeiten der elektronischen Publikation wirklich fast zwei Jahre vergehen müssen, bis die Ergebnisse eines schnell fortschreitenden Gebietes auf Papier vorliegen. Der Rezensent schwankt noch, was ihm mehr Spaß macht: die schnelle und aktuelle Info via Internet oder die bequeme Lektüre der Bücher. Nach der Lektüre kann man noch ein bißchen träumen: Wie wäre es, wenn COSMAS eine (syntaktische) Kontextsuche à la DELIS hätte, wenn es ein standardisiertes Tagset für das Deutsche gäbe, das auch in der von Bochum ins Internet gestellten DUDEN-Version enthalten wäre und so weiter ...

Hypertext – Information Retrieval – Multimedia '97

Theorien, Modelle und Implementierungen integrierter elektronischer Informationssysteme

Universität Dortmund, Lehrstühle Informatik I und VI
29. September – 2. Oktober 1997

Die Fachtagung wurde gemeinsam vom Hochschulverband Informationswissenschaft, den GI-Fachgruppen Hypertext, Information Retrieval und Multimediale elektronische Dokumente, der Östereichischen Computer Gesellschaft sowie der Schweizer Informatiker Gesellschaft veranstaltet.

Gerhard Knorz

1 Drei Fachgruppen – eine Tagung

Kaum eine andere Idee hat wie das World Wide Web neben den Perspektiven auch die Gewohnheiten so vieler Menschen weltweit in so kurzer Zeit grundlegend verändert. In die Phase der dynamischen Entwicklung fällt auch die Zeitspanne zwischen der ersten gemeinsamen Fachtagung der 3 Fachgruppen der Gesellschaft für Informatik, *Hypertext, Information Retrieval* und *Multimediale elektronische Dokumente* unter dem Acronym HIM '95 in Konstanz und nun der zweiten Tagung dieser Reihe in Dortmund. Wie wenig es Sinn macht, elektronische Informationssysteme nur aus *einem* Blickwinkel heraus zu konzipieren, wie stark also eine interdisziplinäre Sicht die richtige Perspektive auf einen Gegenstandsbereich ist, der sich noch nie zuvor so weit an die Bürger einer sich tatsächlich herausbildenden Informationsgesellschaft herangetreten ist, ist 2 Jahre nach einem ersten Experiment zu einem selbstverständlichen Credo aller Beteiligter geworden. Die HIM '97 war angetreten, sowohl für eine theorie- und methodenorientierte Diskussion wie auch für anwendungsorientierte Fragen eine anspruchsvolle Plattform zu bieten.

Die Tagung wurde von den Lehrstühlen 1 und 6 der Universität Dortmund organisiert. Vom Hauptbahnhof Dortmund liegen Universität – und damit auch der Tagungsort – im Prinzip einfach mit der S-Bahn zu erreichen. In den langen Zeiten des Wartens am Bahnhof hatte ich genügend Muße, die besondere Eignung des Tagungsortes für sein Thema wahrzunehmen. Das Interface der Bahn

zu ihren Kunden ist eine Fundgrube für jeden, der den Gründen mißlungener Interaktion und Nutzung nachgehen will: Von einem Fahrplan mit Fußnoten bis zur eigenen Logik der Fahrkartenautomaten und all den anderen Hindernissen auf dem Weg zu einem Platz im Abteil. Und seitdem ich die Unwägbarkeiten der Abfahrtszeiten wiederholt erfahren durfte, habe ich meine eigene Theorie darüber, warum Dortmund in der IR-Szene als das deutsche Mekka probabilistischer Ansätze gilt.

Die Tagung selbst bestand aus einem breitgefächertem Angebot wissenschaftlicher Beiträge auf gutem Niveau (WWW-Anwendungen, Interfaces, Digital Libraries, automatische Methoden der Hypertexterstellung u. a.) und einer begleitenden Ausstellung, z. T. korrespondierend zu einzelnen Vorträgen. *Eine* – die einzige – Enttäuschung darf man nicht beiseite lassen: Daß zu diesem unbestritten aktuellen Thema mit Mühe und Not vielleicht 100 Teilnehmer der Einladung von 3 kooperierenden Fachgruppen nach Dortmund zu locken waren, bedarf einer Erklärung, die ich selbst auch nicht anbieten kann. Statt Spekulationen hier der Hinweis für Daheimgebliebene: Der Tagungsband, herausgegeben von Norbert Fuhr, Gisbert Dittrich und Klaus Tochtermann ist als Band 30 der Schriften zur Informationswissenschaft beim Universitätsverlag Konstanz (UVK) erschienen.

2 Vorträge

Im folgenden werde ich schlaglichtartig auf einige der Vorträge eingehen, wobei die Auswahl von meinem eigenen Interesse, z. T. aber auch von Randbedingungen bestimmt ist, die unabhängig von der Tagung selbst sind (z. B. von der Abfahrwahrscheinlichkeit der S-Bahn).

2.1 Multimedia: Technologien und Strategien im Rundfunkbereich

Der erste Hauptvortrag von L. Danilenko (Technische Direktion des WDR, Köln) geht auf eine fesselnde und vergnügliche Art der Frage nach, welche Zielgruppen der öffentliche Rundfunk mit welchen Strategien bedienen sollte. Die Mediennutzer beschreibt Danilenko als Mischung aus 3 Grundtypen mit völlig unterschiedlichen Anforderungen an und Potential für Medienanbieter:

- Der *homo sapiens* benötigt Informationen als Basis seiner Urteilsfähigkeit. Stichwörter sind Bildung, Wissen, Mitgestaltung. Spartenprogramme können eine der Antworten auf Anforderungen dieses Typs sein.

- Für den *homo economicus* steht der Nutzen im Vordergrund. Für Spielereien hat er keinen Sinn. („Erst das Fressen, dann die Moral", [Brecht]). Eine thematische Antwort der Anbieter geht in Richtung Börse, Umwelt, Gesellschaft oder Testergebnisse. Auf der technischen Seite ist Videotext ein Angebot. Wenn es Vorteile bringt, toleriert der *homo economicus* auch eine komplexe Bedienung.
- Der *homo ludens* baut auf dem ältesten Verhaltenserbe des Menschen auf. Schließlich ist das Spielen der Tiere das überlebenswichtige Äquivalent für die Schule der Menschen. Wenn auch die Kontrolle der Medien beim *homo economicus* angesiedelt ist, erfunden wurden Sie vom *homo ludens*. „Niemals Langeweile" lautet die Anforderung dieses Konsumententyps. Der *homo ludens* akzeptiert auch eine teure Ausrüstung, er ist ein Umsatz-Intensivierer.

Vom gesetzlichen Auftrag an den Öffentlichen Rundfunk her („landesweit und gleichwertig") ergibt sich die Konsequenz eines Vollprogramms, das keinen der Bereiche Information, Bildung und Unterhaltung ausschließt. Die von Danilenko vetretene Strategie des öffentlichen Rundfunks muß die eines integrierten Ansatzes sein: Vernetzen statt Verspartung! Herstellen von Kontext, Bezügen und Zusammenhängen. Unter dem Diktat knapper Kassen geht es zunächst um sparsame Mehrwert-Angebote wie MUX (Hintergrundberichte, neue Sendereihenfolgen) und Festival (wiederholte Nutzung von Eigenproduktionen). Ebenfalls vor der Tür: Filter für die persönliche Programmauswahl.

Im folgenden vergleicht Danilenko Fernseh- und Internet-Nutzung quantitativ und geht auf den Wechsel im Medienverständnis ein, den der Übergang von der Pull- zur Pushtechnologie mit sich bringt, so wie er gegenwärtig von Microsoft und Netscape vorangetrieben wird. Eine lebhafte Diskussion im Anschluß belegt ein intensives Interesse der Tagungsteilnehmer.

2.2 Navigation und Suche – zwei komplementäre Ansätze für multimediales Information Retrieval

Der große Name bei den eingeladenen Vorträgen ist Yves Chiaramella mit seinem Beitrag "Browsing and Querying: Two Complementary Approaches for Multimedia Information Retrieval". Wie kaum ein anderer steht Chiaramella für Arbeiten, die man allen verschiedenen Schwerpunkten der HIM zuordnen kann. Sein Vortrag ist nicht dazu da, revolutionäre Erkenntnisse einem überraschten Auditorium zu präsentieren. Das Verdienst des Beitrags liegt vielmehr in der didaktischen

Aufbereitung der Botschaft, daß die gegenwärtige Kombination von Suche und Navigation bisher über eine Addition nicht hinausgeht und daß sehr wohl ein Bedarf und auch ein Potential für integrierte Ansätze besteht. Der Weg dazu führt weg von der Sicht, die Dokumente als atomare Einheiten betrachtet, hin zu einem flexiblen Dokumentenbegriff, der strukturelle Differenzierungen innerhalb von (Web-)Seiten und über Seitengrenzen hinaus vorsieht.

2.3 Ein Streifzug durch die Tagung

Einen direkten thematischen Bezug zu Chiaramellas Beitrag hatte der Vortrag von Marc Rittberger aufzuweisen: *„Kontextsensitive Visualisierung von Suchergebnissen"* (Bekavac/Rittberger). Hintergrund ist das Konstanzer Projekt *„Virtuelle Informationsräume"* für die Suche in elektronischen Marktplätzen. Es wird ein Verfahren vorgestellt, das Suchtreffer nicht als Liste isolierter Informationseinheiten (Dokumente) präsentiert, sondern das die Treffer in ihrem strukturellen Kontext darstellt.

Daß Retrieval und speziell die Frage nach der effektiven Unterstützung explorativer Suchen eine Problemstellung ist, auf die nicht nur IR-Spezialisten kommen, beweist der Beitrag *„Visualisierung zur Unterstützung der Suche in komplexen Datenbeständen"* von Elzer/Krohn (Institut für Prozeß- und Produktionsleittechnik der TU Clausthal). Anlaß war die unerwartete Schwierigkeit, die Wiederverwendbarkeit von Softwareentwürfen durch Retrievalprozesse in entsprechenden Datenbanken effektiv zu unterstützen. Ein experimentelles Demonstrationssystem, das zur Suche und Ergebnisanzeige einen 3D-Raum verwendet, wurde auf der Basis von Smart (Salton), einem kommerziellen Statistikpaket (SAS) und einer kommerziellen Visualisierungssoftware entwickelt.

Einen ganz andereren Beitrag zur Gestaltung von Retrieval-Interfaces liefern Christian Wolf (Universität Leipzig) und Christa Womser-Hacker: *„Graphisches Faktenretrieval mit vager Anfrageinterpretation"*. Werkstoffinformation kann natürlichsprachig und durch Manipulation von Kurven in Liniengraphiken recherchiert werden. In beiden Fällen bietet Fuzzy Logic die Grundlage der adäquaten Behandlung vager Konzepte.

Mounia Lalmas stellt in sehens- und hörenswerter Weise *"A model for structured document retrieval: empirical investigations*" (Lalmas/Ruthven) vor – ein Beitrag von der Universität Glasgow. Es geht darum, zu entscheiden, ob Dempster-Shafer's Evidenztheorie prinzipiell in der Lage ist, die Aggregation von Relevanzentscheidungen von isolierten Dokumentkomponenten zu modellieren.

In Ermangelung einer wirklich geeigneten Testkollektion basiert die „empirische Untersuchung“ auf einer sehr formal definierten Versuchsanordnung, in der das Ergebnis mehr von dieser Versuchsanordnung als von der untersuchten Evidenztheorie abhängt. Nun ja, bis auf diesen kleinen Schönheitsfehler hatte der Beitrag alles, was man von einem guten Beitrag erwartet.

Deutlich weiterführender erscheint mir der Dortmunder Beitrag *“Probabilistic Logical Information Retrieval for Content, Hypertext, and Database Querying*“ (Rölleke/Blömer), in dem der programmatische Beitrag von Fuhr auf der HIM'95 über ein probabilistisches Datalog („Deduktive Datenbanken“) auf der Basis der Implementierung HySpirit weiterentwickelt und evaluiert wird

“DVS – A System for Recording, Archiving and Retrieval of Digital Video in Security Environments” (Herzner/Kummer/Thuswald) ist in dem Sinne spannend, als der Beitrag die praktisch zu lösenden Probleme von wirklich großen existierenden Anwendungen (Größenordnung von bis zu 1 000 installierten Analog-Videokameras) schildert und eine technische Antwort darauf vorstellt.

Eine Anwendung, deren Nutzen sich mir erst beim zweiten Nachdenken überzeugend erschlossen hat, kommt von der Universität Alcal: “*A User Interface for the Design of Human Figures Multimedia Animations*”. Es geht darum, wie man virtuelle Menschenpuppen effizient zum Tanzen bringt, und in Ermangelung einer breit akzeptierten Notation für Choreographie hat man mit einem entsprechenden System gleich auch ein nachgefragtes Kompositionswerkzeug für Choreographen.

Der Beitrag *„Transformationelle Multimedia-Softwareentwicklung“* (Boles/Wüterich) wird mit viel Engagement vorgetragen. Der Vortragende kritisiert die Dominanz der Informatiker bei der Entwicklung von Multimedia-Produkten und überzeugt dann aber vermutlich nur wenige, daß genau diese Dominanz sich nicht in dem vorgestellten Prinzip der transformationellen Softwareentwicklung wiederfindet.

Für mich einer der besten Beiträge der Tagung kommt von Christoph Baumgarten: *„Probabilistische Modellierung der effizienten Informationssuche in verteilten multimedialen Dokumentbeständen durch Einschränkung des Suchraumes“*. Ich nutze hier den Titel gleich als ein Abstract und beschränke mich auf den Hinweis, daß das behandelte Thema im Kontext von WWW und Digital Libraries keinesfalls nur von theoretischer Bedeutung ist. Ein verwandtes Problem behandelt Norbert Gövert (Universität Dortmund): *„Evaluierung eines entscheidungstheoretischen Modells zur Datenbankselektion“*. Negatives Ergebnis dieser (kleineren) Untersuchung ist die Tatsache, daß die Selektion geeigneter Datenbanken

für ein Suchproblem die Schätzung der Anzahl relevanter Dokumente der Datenbank erfordert, und daß diese Schätzung mit einfachen Mitteln sich als nicht möglich herausstellt.

Ein sehr interessanter Ansatz aus Frankreich verwendet Ansätze von Information Extraction (Eigennamen, Ereignisse, Rollen), um aus Agenturmeldungen automatisch einen zusammenhängenden Hypertext zu konstruieren: *"High Precision Hypertext Navigation based on NLP Automatic Extractions*" (Vichot/Tomeh/Dillet/Wolinski/Guennou/Aydjian). Ebenfalls mit dem Thema „Hypertext und Agenturmeldungen" setzt sich der Beitrag "*Automatic Construction of News Hypertext*" (Dalamagas/Dunlop) von der Universität Glasgow auseinander. Clustertechniken werden eingesetzt, um Links zwischen Dokumenten zu erzeugen. Durch Eliminieren bestimmter Links entstehen unter Berücksichtigung der zeitlichen Reihenfolge der Nachrichten zusammenhängende Stories.

Einen Einblick in die Alltagswelt eines Suchmaschinenbetreibers liefert der Beitrag *„Realisierung und Optimierung der Informationsbeschaffung von Internet-Suchmaschinen am Beispiel von www.crawler.de"*. Mit Hilfe umfangreicher statistischer Erhebungen wird belegt, daß tatsächlich ein Zusammenhang besteht zwischen der Hierarchietiefe eine Dokumentes und der Wahrscheinlichkeit, daß es geändert wird. Fazit für die Suchmaschine: Je kürzer der Pfad eines Dokumentes, desto häufiger muß es auf Änderung überprüft werden.

3 Zusammenfassung

Eine Tagung zu veranstalten, ist für viele beteiligte Personen ein Unternehmen mit vielen zusätzlichen Sorgen und Arbeitsaufträgen. Was das wissenschaftliche Programm der HIM '97 anbelangt, so hat sich dieser Aufwand ganz sicher gelohnt. Auch die Organisation der Tagung lief völlig reibungslos. Ein Kompliment an die Dortmunder Veranstalter! Was an Wünschen bleibt? Man hätte der Tagung noch deutlich mehr Teilnehmer gewünscht. Und für 1999 wünsche ich mir eine Tagung, deren Scope dem der HIM entspricht – mindestens, denn die Designer stehen (zu recht) immer höher im Kurs, aber auf Tagungen wie dieser immer noch außen vor. Aber bis dahin ist ja noch Zeit. Zum Beispiel dafür, den Tagungsband nochmals etwas genauer zu studieren!

„Hermeneutik, Semiotik und Informatik“

Tagung an der Berlin-Brandenburgischen Akademie der Wissenschaften, 28. und 29. November 1997

Winfried Lenders
Universität Bonn
e-mail: lenders@uni-bonn.de

Am 28. und 29. November 1997 fand an der Berlin-Brandenburgischen Akademie der Wissenschaften eine Tagung zum Thema „Hermeneutik, Semiotik und Informatik“ statt, die von Ferdinand Fellmann (TU Chemnitz), Dietmar Roesner (TU Magdeburg) und Jürgen Trabant (FU Berlin) konzipiert und geleitet wurde. Eingeladen waren ca. 30 Wissenschaftler, die man den in der Thematik genannten Gebieten zuordnen kann. Hintergrund der Tagung war die Vorbereitung einer interdisziplinären Arbeitsgruppe an der Akademie, deren Ziel es sein soll, Perspektiven für eine praktische Zusammenarbeit von Hermeneutik, Semiotik und Informatik zu erarbeiten.

Nach ausführlichen Eröffnungsstatements der drei Initiatoren kamen in drei aufeinanderfolgenden Sektionen die Experten der drei Gebiete zu Wort. Um es vorweg zu sagen: Zwar wurden in den Eröffnungsstatements noch die gemeinsamen Fragestellungen genannt, im weiteren Verlauf aber ging es im wesentlichen um Feldbestimmungen und Abgrenzungen, und am Ende herrschte eigentlich Ratlosigkeit darüber, wie denn diese so unterschiedlichen Disziplinen zusammen kommen oder voneinander profitieren könnten.

Gemeinsames Band, so schien es am Anfang, könnte die Frage nach dem Verstehen sein, eine Frage, die, wie Ferdinand Fellmann hervorhob, stets die eigentliche Kernfrage der philosophischen Hermeneutik gewesen sei. In der Hermeneutik dieses Jahrhunderts (Dilthey, Gadamer) sei Verstehen niemals bloß rezeptiv, sondern immer an die Gegenwart, an das Erleben des Verstehenden gebunden. Verstehen heiße danach vor allem auch Explizitmachen des individuellen Bedeutungshintergrundes, des Weltwissens des verstehenden Individuums. Zieht man nun in Betracht, daß Verstehen an Texte - bzw. allgemeiner - an Zeichen gebunden ist, so zeigen sich die Fragestellungen der Semiotik: Wie kann man den Verstehensprozeß systematisch beschreiben, wissenschaftlicher Erforschung zugänglich machen? Welche Regelhaftigkeiten lassen sich etwa beim Verstehen literarischer Texte ausmachen? Wie sind Texte aufgebaut, und was geschieht im Bewußtsein des Menschen, wenn er sie rezipiert? Fragen dieser Art werden, wie

Jürgen Trabant ausführte, heute in der Semiotik, vor allem aber auch in der literaturwissenschaftlichen Hermeneutik erörtert. Von dort aus ist der Weg nicht mehr weit zur Operationalisierung des Verstehensbegriffs, wie er sich in der Informatik findet. Auch hier geht es, wie Dietmar Rösner hervorhob, um Texte, und zwar um das Verstehen und Erzeugen von Texten durch den Computer, um die Fähigkeit des Computers, Fragen zu erkennen und zu beantworten oder Texte in andere Sprachen zu übersetzen. Durch Operationalisierung dessen, was beim Verstehen abläuft, versuche man hier, den Prozeß des Verstehens selbst zu erklären.

Die in dieser Eröffnung debattierten Probleme beherrschten die folgenden Diskussionen. Dabei reichte im Falle der Hermeneutik das Spektrum der Meinungen von der strengen Forderung Georg Meggles, zunächst einmal zu definieren, was man denn mit ‚Verstehen' meine, bis zu der Meinung, daß Verstehen nicht objektivierbar, operationalisierbar sei. Bei aller Einigkeit hierüber stellte sich doch immer wieder die Frage, die Ferdinand Fellmann als Leitfrage formuliert hatte, ob man sich nicht ein Instrument vorstellen könne, das dem Hermeneutiker beim Verstehen eines Textes helfen könne, indem es z. B. die an einen Text zu stellenden Fragen herausfinde und die dazu passenden Antworten generiere. Ein solches Konzept, so führte Burkhard Liebsch aus, reduziere das Verstehen auf Informationsfragen, und so könne man sich einen kreativen Umgang mit Texten nicht vorstellen. Ein Ausweg könnte vielleicht darin bestehen, den Verstehensbegriff in Richtung auf die Informatik „tieferzuhängen", wie Thomas Rolf meinte, doch stellt sich dann die Frage, ob man damit nicht im Grunde das Anliegen der Hermeneutik aufgibt.

Stand unter den Hermeneutikern noch der Verstehensbegriff als gemeinsames Band im Vordergrund des Interesses, so ging es den anschließend vortragenden Semiotikern in erster Linie darum zu klären, um welche Probleme es in der Semiotik geht. So stellte Achim Eschbach zunächst einmal klar, was Semiotik heute überhaupt ist: Semiotik sei nicht eine statistische Disziplin, die sich auf den Zeichencharakter richte, sondern verstehe sich als Deutungstheorie des Zeichens. Es gehe ihr um die kommunikativen Ereignisse als grundlegende soziale Aktivitäten. Im weiteren Verlauf der Diskussion spielte dann neben literaturwissenschaftlichen Sichtweisen vor allem die Semiotik von Charles Sanders Peirce eine herausragende Rolle. Nach Peirce sei, wie Susanne Rohr ausführte, mit jeder Zeichenrelation eine individuelle Perspektivierungsleistung verbunden, die sich nicht nur auf das Subjekt, sondern auch auf das Objekt richte und daher objektkonstituierend sei. Zeichenobjekte seien damit kulturelle Produkte und individuelle Interpretation. Verstehen von Zeichen sei daher immer auch ein Prozeß des interpretie-

renden Erschaffens. Vor allem diese Sichtweise sei es, die von der Semiotik in die Diskussion mit der Hermeneutik und Informatik eingebracht werden könne.

Nach dieser Beschreibung der Hermeneutik und Semiotik war die Erwartung groß, welche Anknüpfungspunkte es denn nun in Richtung Informatik gebe. Ebenso groß wie die Erwartung war auch die Ernüchterung. Denn sowohl Egbert Lehmann als auch Herbert Stoyan stellten klar, daß Informatik es immer nur mit Berechenbarem, Formalisierbarem zu tun habe. Wenn man mit Methoden der Informatik an eine Frage herangehe, müsse man sich, so Stoyan, die Frage stellen, was bei der Formalisierung ablaufe, was überhaupt formalisierbar sei und wie man formalisieren solle. Beim Formalisieren werde alles entfernt, was der Hermeneutik wichtig sei (deutlicher geht es wohl nicht mehr!). Formalisierbar sei nur, was semantisch klar sei, was auf einfache Art formulierbar sei und was sich aus Wissenskontexten lösen lasse. Diese deutlichen Aussagen lösten beim Publikum einerseits Zerknirschung (Trabant), andererseits Humor aus: gehe es doch der Philosophie immer schon um das kontextfreie Ding an sich (Fellmann). Doch die Rettung war nicht weit: Bei aller Vorsicht dürfe man aber auch nicht davor zurückweichen, die Möglichkeiten zu explorieren, die man vielleicht doch habe (Lehmann). So gebe es z. B. durchaus Erfolge in der Untersuchung und maschinellen Simulation des Lernen aus Beispielen und aus Erfahrungen, in der Maschinellen Übersetzung und bei der Entwicklung von natürlich-sprachlichen Systemen. Und weiterhin, so Lenders, habe Stoyan durchaus der Hermeneutik einen Weg gezeigt, wie sie von der Informatik profitieren könne: Man müsse dazu jedoch aufzeigen, was klar und formalisierbar sei. Nichts anderes habe Winograd 1972 gemacht; auch im System LILOG, das von Manfred Stede vorgestellt wurde, seien bestimmte Teile des natürlich-sprachlichen Verstehensprozesses modelliert worden, soweit diese sich klar und formalisierbar beschreiben ließen. Nicht anders könne es gehen: Die Hermeneutik müsse beschreiben, was bei ihr klar sei und einfach, müsse vielleicht auch ein Kontextmodell entwickeln und definieren, was Verstehen eigentlich sei. Wenn man Tools zur Unterstützung hermeneutischer Arbeit wolle, müsse man beschreiben, was diese im einzelnen leisten sollten.

Schließlich ging man auseinander in dem Bewußtsein, die jeweiligen Ansätze und Ansichten besser kennengelernt zu haben, und mit der Absicht, den Dialog fortzusetzen. Um das nächste Gespräch zu konkretisieren, sollen alle, Hermeneutiker, Semiotiker und Informatiker, ihren jeweiligen *state of the art* in Sachen Textbearbeitung am Beispiel vorführen und mögliche Erwartungen an die anderen formulieren.

Strukturen und Relationen in der Wissensorganisation

Die International Society for Knowledge Organization (ISKO) wird vom 25. bis 29. August ihre 5. internationale Konferenz in Lille (Frankreich) unter dem Titel **Structures and Relations in Knowledge Organization** veranstalten. Aus dem Call for Papers entnimmt man folgendes:

> The conference will focus on the role of relationships and emergent knowledge structures as represented in the human mind, in information handling tools (including classification schemes, thesauri, and indexing systems) and in computers and intelligent/knowledge-based systems.
>
> Papers and panels address the conference theme from any of the following perspectives:
>
> 1) **Theory of knowledge organization:** History, paradigms, philosophy, societal aspects, epistemology, division of the sciences.
> 2) **Disciplinary and interdisciplinary approaches to knowledge organization:** Formalization of structures and relations in and across linguistics, semiotics, cognitive sciences, computer science, artificial intelligence, etc.
> 3) **Cognitive approaches to knowledge organization:** Conceptual entities and interconcept relations, category formation, classical and non-classical classifications and their use in information organization and retrieval, concept representation in knowledge-based systems, object-oriented analysis and design, types of relations.
> 4) **Design of information systems:** Structure and relations in indexing and retrieval languages, design of controlled vocabularies, terminology building and extraction tools, thesauri and metathesauri, multilingual thesauri, standardization of relationships, problems of compatibility.
> 5) **The comparative approach:** Common and particular relationships in different knowledge systems.
> 6) **Linguistics in knowledge organization:** Structure and relations in sublanguages/special purpose languages/technical writing, discourse structures and relations, intelligent text processing, natural language

processing-based systems and their use in knowledge representation and extraction.

7) **New technologies for knowledge organization:** Structures and relations in the online environment, applications of classical and non-classical structures to computer-based indexing and retrieval systems, search engines, distributed and multilingual knowledge bases.
8) **Conceptual modeling :** Data modeling, knowledge modeling, user profile modeling.
9) **Universals of structures and relations** in knowledge organization.

Contact: Widad Mustafa Elhadi (conference chair), UFR IDIST, University Charles de Gaulle Lille 3, BP 149, 59653 Villeneuve d'Ascq, France. E-mail: isko.conf@univ-lille2.fr .

BOBCATSSS '99: Learning Society – Learning Organisation – Lifelong Learning

25th – 27th of January 1999 – Bratislava

BOBCATSSS '99 is the seventh BOBCATSSS-symposium. This time it takes place in Bratislava. It will be attended by all kinds of information professionals in companies and institutions (e.g. information managers, librarians, lecturers and students).

Organisation: The BOBCATSSS-symposium is arranged by a students' project of the Fachhochschule Darmstadt – University of Applied Science, Department of Information Science and the University of Library and Information Science Stuttgart. In Bratislava the local organiser is the Department of Library and Information Science at the Comenius University. The organising students define this project as an example of a learning organisation and self-directed learning. The Proceedings will be published at the time of the conference.

Topic: BOBCATSSS '99 aims to point out what kind of changes and challenges make a learning society necessary.

- New information and communication technologies have introduced a change towards an information society
- Free, world-wide transfers of goods, capital, and services have led to a globalization of markets
- There is a notable increase of polarisation within and between societies

Future developments of firms and institutions depend on their ability and will to implement collective learning processes and a change of organisation processes as well as their ability to change into a learning organisation.

Collective learning means using organisations' internal knowledge and capacity as well as their further development by a constant systematic achievement of new explicit knowledge. Creative solutions and innovations can only be achieved if organised learning takes place in altered organisation structures:

- reduction of hierarchies
- increase in teamwork
- empowerment of the workforce

- using information technologies
- systematically controlled information processes.

Within the learning society citizens need support to solve the problems of the outlined changes in all fields of life, especially in working life. As we face the ever faster obsolescence of knowledge, job prospects by means of lifelong learning need to be ensured. New models and possibilities which enable self-directed learning are required.

New educational media and new types of distance learning developing from information technologies are important for self-directed learning independent of place, time, and teachers. Papers will cover all aspects of LLL, especially on the following topics:

- The crucial role of information technologies for learning organisations and lifelong learning
- Information management and knowledge management in learning organisations
- Self-directed learning in organisations – the role of information managers
- Libraries – learning organisations and agencies for lifelong learning

Registration: For more information and registration please contact: http://www.fh-darmstadt.de/BOBCATSSS/conf99.htm .

KONVENS 98
Computer, Linguistik und Phonetik zwischen Sprache und Sprechen

4. Konferenz zur Verarbeitung natürlicher Sprache

5.-7. Oktober 1998, Universität Bonn

Die KONVENS ist die zentrale Tagung aller wissenschaftlichen Gesellschaften im deutschsprachigen Raum, die sich mit Sprache und Computern befassen. Sie findet im 2-jährigen Turnus statt. 1998 wird die KONVENS federführend von der Gesellschaft für Linguistische Datenverarbeitung (GLDV) in Bonn ausgerichtet, zusammen mit der Deutschen Gesellschaft für Sprachwissenschaft (DGfS), der Gesellschaft für Informatik (GI), FA 1.3 „Natürliche Sprache", der Informationstechnischen Gesellschaft/Deutschen Gesellschaft für Akustik (ITG/DEGA) und der Österreichischen Gesellschaft für Artificial Intelligence (ÖGAI).

Thema der Tagung sind alle Bereiche der maschinellen Verarbeitung von Sprache in geschriebener und gesprochener Form. Besondere Aufmerksamkeit soll dabei solchen Ansätzen zuteil werden, die sich mit den strukturellen und phonologisch/phonetischen Aspekten der computerunterstützten Sprachforschung befassen und dazu beitragen, eine Brücke zwischen diesen beiden Aspekten zu schlagen.

Das Tagungsprogramm wird Einzelvorträge, Workshops, Systemvorführungen und Postervorführungen umfassen. Tagungssprachen sind Deutsch und Englisch. Der Tagungsband wird rechtzeitig zu Beginn der Tagung gedruckt vorliegen. Die KONVENS 98 findet im Hauptgebäude der Universität (Stadtzentrum, in Gehnähe zum Hauptbahnhof) Bonn statt.

Die örtliche Organisation liegt bei Prof. Dr. Wolfgang Hess, Prof. Dr. Winfried Lenders, Dr. Thomas Portele und Dr. Bernhard Schröder.

Das Programmkomitee besteht aus Dr. Ernst Buchberger, Wien (ÖGAI), Dr. Stefan Busemann, Saarbrücken (GI), Prof. Dr. Dafydd Gibbon, Bielefeld (DGfS), Prof. Dr. Roland Hausser, Erlangen (GLDV), Prof. Dr. Wolfgang Hess, Bonn (ITG/DEGA), Prof. Dr. Wolfgang Hoeppner, Duisburg (GI), Prof. Dr. R. Hoffmann, Dresden (ITG/DEGA), Dr. Tibor Kiss, Heidelberg (DGfS), Prof. Dr. Winfried Lenders, Bonn (GLDV), Dr. Harald Trost (ÖGAI).

Kontakt: Das Tagungsbüro ist am Institut für Kommunikationsforschung und Phonetik der Universität Bonn eingerichtet:

- Adresse: Poppelsdorfer Allee 47, D-53 115 Bonn
- Aktuelle Information im Internet:
  http://www.ikp.uni-bonn.de/Konvens98
- e-mail: konvens98@uni-bonn.de
- Tel.: +49-228-73 5638
- Fax: +49-228-73 5639

INFORMATIK '98
Informatik zwischen Bild und Sprache

28. Jahrestagung der Gesellschaft für Informatik

21. – 25. September 1998 – Otto-von-Guericke-Universität Magdeburg

Die Tagung Informatik '98 findet in der Zeit vom 21. – 25. September 1998 in Magdeburg statt und wendet sich an Interessierte aus Wirtschaft, Wissenschaft, Ausbildung und Politik. Die führende Informatiktagung im deutschsprachigen Raum steht 1998 unter dem Motto „Informatik zwischen Bild und Sprache".

Gegenstand der Tagung sollen daher unter anderem alle Aspekte der Generierung, Analyse und Verarbeitung von bildhaften und sprachlichen Elementen in der Informatik sowie deren Anwendung in Technik, Medizin, Natur- und Geisteswissenschaften sein. Während der Tagung wird außerdem eine Halbtagsveranstaltung zu modernen Studiengängen in der Informatik durchgeführt.

Umrahmung der eigentlichen Tagung bilden Workshops, Tutorien, ein Studierendenprogramm und ein Ehemaligentreffen von Absolventen der Universität Magdeburg. Workshops zur Ergänzung und Vertiefung des Tagungsthemas finden im Umfeld der Tagung statt. Sie werden in der Regel in Kooperation mit speziellen Fachgruppen der Gesellschaft für Informatik organisiert. Vorgesehen sind unter anderem:

- Abstract State Machines, Peter Schmitt, Uwe Glässer
- Data Mining and Data Warehousing als Grundlage moderner entscheidungsunterstützender Systeme, Rudolf Kruse, Gunter Saake
- Frauen und Männer in der Informationsgesellschaft, Gabriele Winker
- Informatikanwendung in afrikanischen Ländern, Nazir Peroz
- Integration heterogener Softwaresysteme, Stefan Conrad
- Molekulare Bioinformatik, Ralf Hofestädt
- Multimedia-Datenhanken und -Informationssysteme, Klaus Meyer-Wegener
- Multimedia-Systeme für Wissenschaft und Technik, Hans-Jürgen Appelrath
- Personal Computing im WEB, Werner Remmele, Konrad Klöckner
- Sportinformatik, Heinz Bayen, Jürgen Perl

- Sprachtechnologie: Aufgaben und Herausforderungen für die Informatik, Dietmar Rösner

Zu einem „Computer Animation Festival“, welches als Abendveranstaltung vorgesehen ist, können Beiträge, inbesondere einschlägige Videoproduktionen, Computeranimationen, Computergraphiken u.ä. eingereicht werden. Die besten Arbeiten werden prämiert (Ausschreibung siehe WWW-Seite).

Veranstalter: Gesellschaft für Informatik e. V., Bonn; Otto-von-Guericke-Universität Magdeburg, Fakultät für Informatik.

Tagungsleitung: Prof. Dr. Jürgen Dassow (Universität Magdeburg).

Kontakt: Tagungssekretariat Informatik '98, Fakultät für Informatik, Otto-von-Guericke-Universität Magdeburg, Postfach 4120, D-39 016 Magdeburg, Tel.: +49-391/67-18718, Fax: +49-391/6 12018, Email: gi98@cs.uni-magdeburg.de . Informationen auf dem neuesten Stand: http://www.cs.uni-magdeburg.de/gi98 . Für die Aufnahme in den Informationsverteiler „Informatik '98“ ist das Tagungssekretariat zuständig.

ISI '98 – Knowledge Management und Kommunikationssysteme

6. Internationales Symposium für Informationswissenschaft

4. bis 7. November 1998 – Karls-Universität zu Prag

Nach Konstanz, Oberhof, Saarbrücken, Graz und Berlin wird das 6. Internationale Symposium für Informationswissenschaft vom 4. bis 7.11.1998 in **Prag** stattfinden. Im Rahmen der 750-Jahr-Feiern zum Bestehen der Stadt wird die Tagung vom Institut für Informations- und Bibliothekswesen (IISL) der Karls-Universität zu Prag gemeinsam mit dem Hochschulverband Informationswissenschaft (HI), vertreten durch die Fachrichtung Informationswissenschaft (IW) an der Universität des Saarlandes zu Saarbrücken, veranstaltet.

Knowledge Management und Kommunikationssysteme. Mit diesem gewählten Leitthema trägt auch die diesjährige Tagung der anhaltenden Tendenz zu einer stärkeren Marktausrichtung Rechnung. Im einzelnen wurde um Beiträge zu folgenden Teilthemen gebeten: Workflow Management, Multimedia, Knowledge Transfer, Electronic Publishing, The Internet as Knowledge Base, Recent Developments of Information Systems and Tools, The Information Society. Die Tagung ISI '98 wird bisher nicht publizierte Beiträge aus aktuellen Forschungs-, Entwicklungs- oder Anwendungsbereichen präsentieren. Zum Programm werden gleichermaßen empirische, experimentelle und theoretische Arbeiten gehören. Tutorien, Panels und Posters zu aktuellen Themen werden das Tagungsprogramm abrunden. Die informationswissenschaftlichen Ausbildungsinstitutionen sind eingeladen, die besten studentischen Arbeiten seit ISI '96 (Haus-, Seminar- und Diplomarbeiten) für den „Best Student Paper Award" einzureichen.

Wissenschafliche Leitung: Prof. Dr. Jirí Cejpek, Prof. Dr. Harald Zimmermann. Panel: Prof. Dr. Rainer Kuhlen, Poster: Prof. Dr. Gerhard Knorz, Best student paper award: Prof. Dr. Wolf Rauch.

Programmkomittee: Prof. Dr. Harald Zimmermann, Uni Saarbrücken (Vorsitz) Dr. Rolf Assfalg, Uni Konstanz Prof. Dr. Raffael Capurro, FH Stuttgart Prof. Dr. Jirí Cejpek, Karls-Universität Prag Prof. Dr. Hans-Peter Frei, UBS Zürich Prof. Dr. Norbert Fuhr, Uni Dortmund Dr. Hans Giessen, Uni Saarbrücken Prof. Dr. Rainer Hammwöhner, Uni Regensburg Dr. Ilse Harms, Uni Saarbrücken Prof. Dr. Ralf-Dirk Hennings, FH Potsdam Prof. Dr. Norbert Henrichs, Uni Düsseldorf Matthias Herfurth, IZ Bonn Dr. Josef Herget, EMS Konstanz Stephan Holländer, HTL Chur Dr. Sta-

nislav Kalkus, Karls-Universität Prag Prof. Dr. Gerhard Knorz, FH Darmstadt Prof. Dr. Alfred Kobsa, GMD FIT St.Augustin Prof. Dr. Jürgen Krause, IZ Bonn Prof. Dr. Rainer Kuhlen, Uni Konstanz Dr. Heinz-Dirk Luckhardt, Uni Saarbrücken Prof. Dr. Achim Oßwald, FH Köln Dr. Jiri Panyr, Siemens AG München Peter Poschadel, Uni Saarbrücken (stud. Vertreter) Prof. Dr. Wolf Rauch, Uni Graz PD Dr. Ulrich Reimer, Swiss Life Zürich Prof. Dr. Harald Reiterer, Uni Konstanz Dr. Marc Rittberger, Uni Konstanz Dr. Christian Schlögl, FH Eisenstadt Prof. Dr. Ralf Schmidt, FH Hamburg Prof. Dr. Eric Schoop, TU Dresden Prof. Dr. Thomas Seeger, FH Darmstadt Thomas Tanzer, ETH Lausanne Dr. Stephanie Teufel, Uni Zürich Dr. Ulrich Thiel, GMD IPSI, Darmstadt PD Dr. Rudolf Vlasak, Karls-Universität Prag Prof. Dr. Gernot Wersig, Freie Uni Berlin PD Dr. Christa Womser-Hacker, Uni Konstanz.

Kontakt, Tagungsort und -organisation: Institut für Informations- und Bibliothekswesen (IISL) der Karls-Universität zu Prag. Organisation, Auskünfte und Anmeldung (Unterkünfte, Tagungsgebühren „östl.Länder"): ISI '98 c/o Institut für Informations- und Bibliothekswesen (IISL), Karls-Universität, Celetná 20, CZ-110 00 Praha l; Tel: +420-2-24-491520, Fax: -812166. Kontakt: Prof. Dr. Jirí Cejpek, e-mail: jiri.cejpek@ff.cuni.cz. Organisation, Auskünfte und Anmeldung (Abstracts, Tagungsgebühren „westl.Länder") ISI '98 c/o FR 5.5 Informationswissenschaft, Universität des Saarlandes, D-66041 Saarbrücken; Tel: +49-681-302-3537, Fax: -3557; Kontakt: Dr. Volker Schramm, Tel: +49-681-302-3539, -3537; e-mail: v.schramm@is.uni-sb.de.

Herbstschule „Information Retrieval“

1. Herbstschule der Fachgruppe Information Retrieval der Gesellschaft für Informatik

27.September bis 2.Oktober 1998 in Schwerte

Effektive Informationssuche wird immer wichtiger. Deshalb veranstaltet die Fachgruppe Information Retrieval der Gesellschaft für Informatik vom 27.September bis zum 2.Oktober 1998 in Schwerte die erste Herbstschule Information Retrieval. Angesprochen sind alle Interessierten aus Praxis und Wissenschaft, die Nutzer von Suchdiensten und Datenbanken ebenso wie die Entwickler von Informationssystemen.

Ziel der Herbstschule ist die Vermittlung von Kenntnissen über eine gezielte und erfolgreiche Informationssuche. Themen wie Suchverfahren im WWW, Retrieval von Patentinformation, Suchen in multilingualen und multimedialen Daten oder intelligentes und linguistisches Retrieval werden dabei ebenso zur Sprache kommen wie Benutzungsschnittstellen und Verfahren zur Qualitätsbeurteilung.

Weitere Informationen und Anmeldeformulare finden Sie unter http://www.inf-wiss.uni-konstanz.de/IR/. Ansprechpartner: Marc Rittberger 07531/883595 email: ir@inf-wiss.uni-konstanz.de.

Die Heidelberger Akademie der Wissenschaften, die
Gesellschaft für Linguistische Datenverarbeitung und das
Institut für deutsche Sprache *laden ein zum Symposium*

„Computergestützte Produktion und Publikation von Wörterbüchern“

vom 23.9.98 bis zum 25.9.98 in die
Heidelberger Akademie der Wissenschaften

Das Thema:

Seit längerem gilt der Computer als Werkzeug, das die Produktion von Wörterbüchern beschleunigt und verbessert. Hypertext und Multimedia machen den Computer nun auch zum Medium, das Sprach- und Sachwissen auf neuartige Weise darstellen und abrufbar machen kann. Viele Wörterbuchprojekte möchten, unter ihren jeweiligen Rahmenbedingungen, die Möglichkeiten moderner Computertechnik nutzen, um schneller bessere und innovative Produkte herzustellen. Da die technischen Möglichkeiten, die Produktion und Publikation von Wörterbüchern zu unterstützen, sich rasch verändern, ist es jedoch oft nicht leicht, aus den angebotenen Werkzeugen dasjenige auszuwählen, das am besten für das geplante Projekt geeignet ist. Mit dem Symposium möchten wir den Dialog zwischen Anbietern von solchen Werkzeugen und lexikographischen Projekten fördern und möglichst viele an der Thematik Interessierte zusammenbringen, um sich kennenzulernen, Erfahrungen und Ideen auszutauschen und voneinander zu lernen.

Das Programm

Das Programm wird sich aus verschiedenen Bausteinen zusammensetzen:

- Die vormittags stattfindenden Schulungsteile führen in grundlegende Konzepte wie SGML, Datenbanken oder Unicode ein, erörtern den aktuellen Stand von Fragestellungen wie Abrechnungsmöglichkeiten, Archivierung und Urheberrecht im Internet.
- Das Forum „Werkzeuge“ lädt kommerzielle und akademische Software-Entwickler dazu ein, Werkzeuge zur computergestützten Produktion und

Publikation von Wörterbüchern zu präsentieren.

- Das Forum „Projekte“ lädt geplante, laufende und abgeschlossene Wörterbuchprojekte dazu ein, sich vorzustellen und erste Arbeitsergebnisse vorzuführen.
- In den Foren werden Werkzeuge und Projekte in ca. 10-minütigen Kurzpräsentationen vorgestellt; anschließend können sich die Interessierten anhand von Postern und Systemdemonstrationen näher informieren.
- Die Vorträge des letzten Tages behandeln die Perspektiven, die sich durch das Medium Computer für die künftige Wörterbucharbeit ergeben. Zum Abschluß haben wir Expertinnen und Experten dazu eingeladen, mit den Teilnehmern des Symposiums über ihre Vorstellungen und Visionen von der Zukunft der Lexikographie zu diskutieren.

Das Symposium wird organisiert von

- Ingrid Lemberg (Heidelberger Akademie der Wissenschaften, Arbeitsstelle Deutsches Rechtswörterbuch)
- Bernhard Schröder (GLDV)
- Angelika Storrer (IDS)

Die Teilnahmegebühr beträgt 50 DM, für Studierende und Mitarbeiter der Heidelberger Akademie der Wissenschaften, der GLDV und des IDS 30 DM. Die Gebühren sollten an der Tageskasse entrichtet werden.

Die Teilnehmerzahl ist begrenzt, daher empfehlen wir eine zeitige Anmeldung. Das Formular für Ihre Anmeldung können Sie bei Ingrid Lemberg per e-mail (lemberg@drw.adw.uni-heidelberg.de) oder snail-mail anfordern:

Ingrid Lemberg, Akademie der Wissenschaften
Deutsches Rechtswörterbuch
Karlsstr. 4
69 117 Heidelberg

Ein detailliertes Programm findet sich unter http://www.ids-mannheim.de/grammis/program.html. Informationen zum Veranstaltungsort finden sich unter: http://www.ids-mannheim.de/grammis/reise.html.

www.gldv.org – Die Web-Site der GLDV

Bernhard Schröder, Hans-Christian Schmitz

Es sollte ein Leichtes sein – so möchte man meinen – die elektronischen Internet-basierten Formen der Kommunikation effizient in einer wissenschaftlichen Gesellschaft anzuwenden, die Computerlinguistik und Sprachtechnologie zum Gegenstand hat; denn hier darf man natürlich bei der weit überwiegenden Anzahl der Mitglieder die notwendige Infrastruktur und die Bereitschaft, diese Medien als Kommunikationsform in der Gesellschaft anzunehmen, erwarten. Und Reaktionen von GLDV-Mitgliedern bestätigen mir auch, dass diese Erwartung berechtigt sind. Aber der schon professionell bedingten Offenheit der GLDV-Mitglieder für die elektronische Kommunikation stand bei dem Versuch, dies stärker nutzbar zu machen, zunächst die schlechte Datenlage gegenüber. Ein Blick in die Mitgliederdatenbank erinnerte daran, dass der Siegeszug des Internet, gemessen an der Aktualisierungsfrequenz der Datenbank, noch jung ist. Zur Aktualisierung eines Mitgliederdatensatzes kam es ja in der Regel nur bei einem Wechsel der beruflichen Wirkungsstätte oder des privaten Wohnsitzes. Einem Umzug im (oder auch einem Neueinzug in den) Cyberspace wurde in der Regel im Hinblick auf die GLDV keine größere Bedeutung beigemessen. Nur bei etwa einem Viertel der GLDV-Mitglieder war eine e-mail-Adresse überhaupt vermerkt, und darunter war auch mancher längst stillgelegte Bitnet-Knoten vertreten.

Aktualisierung der Mitgliederdaten

Es galt also zunächst die Mitgliederdaten zu aktualisieren, um die Adressen der Mitglieder im virtuellen Raum ausfindig zu machen. Im GLDV-Vorstand kam man schnell überein, dass die Gelegenheit zu einer generellen Revision des Datenbestands genutzt werden solle. Neben den üblichen persönlichen Daten wurden auch Fragen nach Interessensgebieten und Zugehörigkeit zu Arbeitskreisen in den Fragenkatalog aufgenommen. Die letztgenannten Angaben sollen den Arbeitskreisleitern zur Überprüfung ihrer eigenen Datenbestände und können darüber hinaus zur Etablierung des Informationsflusses innerhalb der Arbeitskreise dienen. Aber auch schon die reinen Interessens- und Zugehörigkeitsbekundungen von GLDV-Mitgliedern bezüglich der AKs können eine wichtige Rückmeldung für die Leiterinnen und Leiter von bislang selten oder in stark wechselnden Konstellationen tagenden Arbeitskreisen darstellen.

Als unter den gegebenen Umständen effizienteste Form der Befragung, die eine größtmögliche und zuverlässige Abdeckung versprach, erschien der telefonische Weg. Herr Kurt Thomas (IKP, Universität Bonn) erklärte sich dankenswerterweise bereit, diese Aufgabe zu übernehmen.

Es wäre selbstverständlich unsinnig, ein derart aufwendiges Verfahren alle paar Jahre wiederholen zu wollen. Doch halte ich die Hoffnung für begründet, dass der elektronische Informationsaustausch, wenn er erst einmal hinreichend innerhalb der GLDV institutionalisiert ist, Anreiz genug ist, zumindest die elektronische Erreichbarkeit zu gewährleisten (wie ja auch das Ausbleiben des LDV-Forums manches Mitglied an eine überfällige Änderung der postalischen Adresse erinnert hat). Erleichert werden soll die Aktualisierung der Angaben durch ein WWW-Formular, das z. Zt. unter http://www.gldv.org/aufnahme.htm abrufbar ist.

Dieses Formular wird derzeit noch gleichzeitig für Aufnahmeanträge und Änderungen verwendet. Wünschenswert wäre natürlich, dass die GLDV-Mitglieder die sie betreffenden Informationen in der Mitgliederdatenbank selbst abrufen und einsehen könnten. Mancher Datenbestand veraltet ja gerade deswegen so schnell, weil die betroffenen Personen und Institutionen nicht (mehr) wissen, welche Angaben über sie genau vorgehalten werden. Es versteht sich, dass die mitgliederbezogenen Daten nicht von unautorisierten Personen abrufbar sein dürfen; sie müssen also durch mitgliederspezifische Kennwortvergabe oder Verschlüsselung gesichert sein. Durch das im Beitrag von Bettina Mielke und Christian Wolff in diesem Heft vorgestellte Konzept kryptographiebasierter Kommunikationsformen ergäbe sich auch für das Problem der Bearbeitung personenbezogener Daten über das Internet eine gut in ein Gesamtkonzept elektronischer Kommunikation innerhalb der GLDV integrierbare Lösung.

Die Liste der bekannten e-mail-Adressen von GLDV-Mitgliedern wurde in einen e-mail-Verteiler eingespeist, der inzwischen unter der Adresse GLDV-Mitglieder@sunbsh.ikp.uni-bonn.de zu erreichen ist. An diese Adresse versandte Nachrichten erreichen alle GLDV-Mitglieder, deren e-mail-Adresse in der Mitgliederdatenbank gespeichert ist. Bislang ist dieser Verteiler unmoderiert, d. h. die eingehenden Nachrichten werden automatisch an die Mitglieder weitergeschickt. Bis auf ein anfängliches technisches Problem, das zur Verbreitung von Fehlermeldungen über den Verteiler führte, hat sich dies auch bewährt. Sollte die Adresse sich allerdings zum Anziehungspunkt für unerwünschte Mail entwickeln, muss natürlich zu einer moderierten Liste übergegangen werden. Über weitere bestehende e-mail-Verteiler der GLDV gibt http://www.gldv.org/verteiler.htm Auskunft. Wünsche nach weiteren Verteilern, z. B. für einzelne Arbeitskreise, nehmen wir gerne entgegen.

Die GLDV-Domäne

An der zuletzt genannten WWW-Adresse sieht man schon die wichtigste formale Neuerung der WWW-Präsentation der GLDV: Die Gesellschaft verfügt über eine eigene Domäne, und zwar unter der Top-Level-Domäne **org**. Damit gesellt sich die GLDV im Internet zu anderen nicht-kommerziellen Einrichtungen. Die GLDV ist auf diese Weise neben dem vollständigen URL http://www.gldv.org mit den meisten Browsern auch unter www.gldv.org oder einfach gldv.org erreichbar. Einem freundlichen Sponsoring durch die Bonner Firma **tops.net** (vgl. http://www.topsnet.de), die vielfältig im Bereich von Internet-Dienstleistungen tätig ist, und besonders auch dem Engagement der tops.net-Geschäftsführerin Frau Lucie Prinz verdanken wir, dass für die Gesellschaft durch die Einrichtung und Wartung dieser Domäne keinerlei Kosten entstehen.

Eine eigene Domäne ist abgesehen von Fragen des „Prestiges" und der mnemonischen Bedeutung auch bei einem evtl. Umzug an einen anderen Standort von Vorteil: Die GLDV kann dann ihre gewohnte symbolische Adresse im Internet behalten, auch wenn sich dann „subsymbolisch" alles ändert. Physisch ist der Web-Server derzeit im Institut für Kommunikationsforschung und Phonetik der Universität Bonn zu finden. Hier wird auch die Einrichtung und Pflege der Web-Seiten betreut.

Das „Einrichtungsdesign" der GLDV-Website orientiert sich noch weitestgehend am ersten Internet-Auftritt der GLDV, den Herr Harald Elsen vor ungefähr zwei Jahren mit Engagement auf den Weg gebracht hatte. Eine gewisse Umgestaltung des Designs wird sicherlich mit der Neuauflage des Informationsfaltblattes anstehen: Die GLDV soll sich in den verschiedenen Medien ja mit ähnlichen Wiedererkennungsmerkmalen darstellen; doch sollen die Web-Seiten der GLDV primär informationsorientierte Seiten bleiben, die den Zugriff auf die Seiten auch über eine Telefonmodemverbindung zum Internet nicht zum Geduldsspiel werden lässt.

Das WWW-Angebot der GLDV

Inhaltlich sollen die Seiten zum einen der Repräsentation der GLDV nach außen dienen, aber gleichzeitig auch die GLDV-Mitglieder aktuell informieren. Neben den unerlässlichen Selbstdarstellungen ist sicherlich die beste Außenwirkung dadurch gegeben, dass man Interessenten Einblick in die Aktivitäten der Gesellschaft gibt. Ein möglichst großer Bereich des Internet-Angebots der GLDV sollte deshalb auch für Nichtmitglieder offen bleiben. Das schließt keineswegs aus,

dass gewisse Ressourcen, die z. B. durch Arbeitskreise zur Verfügung gestellt werden, ausschließlich GLDV-Mitgliedern vorbehalten werden.

Welche Angebote soll man auf den GLDV-Seiten erwarten können? Zum einen sollen die Seiten die Fragen nach Anschriften, Vorstands- und Beiratsmitgliedern, AK-Leiterinnen und -Leitern usw. zuverlässig beantworten. Sie sollen für die schon erwähnten Änderungswünsche bei Einträgen in der Mitgliederdatenbank eine geeignete Benutzeroberfäche bereitstellen. Aber natürlich darf sich das Angebot keineswegs im Institutionellen erschöpfen. Insbesondere müssen die Seiten verlässliche und aktuelle Auskunft über laufende und bevorstehende Aktivitäten innerhalb der GLDV bieten.

Ein großer Teil der Aktivitäten liegt in den Händen der Arbeitskreise; deshalb nehmen die Arbeitskreise breiten Raum in der WWW-Darstellung der GLDV ein. Die Arbeitskreise stellen sich inzwischen jeweils mit einer Kurzdarstellung und einer eigenen Seite im Netz vor. Mit den AK-Seiten ist beabsichtigt, möglichst auch immer aktualisierte Informationen zu den AK-Aktivitäten zu bieten; wir hoffen zu diesem Zweck auf eine entsprechende Kooperation mit den Vertreterinnen und Vertretern der AKs. Soweit diese Informationen auch außerhalb der jeweilgen AKs von Interesse sein könnten, werden wir sie nach entsprechender Mitteilung auch auf den allgemeinen Seiten der GLDV bekannt geben.

Ein wesentlicher Faktor für die Informativität des GLDV-Web-Servers ist – ganz in der „Natur“ des Netzes liegend –, welche Verbindungen er zur Außenwelt unterhält. Hier dürfen Links auf andere Einrichtungen nicht fehlen. Aber wichtig sind gerade auch aktuelle Hinweise auf Veranstaltungen, die für die in der GLDV vertretenen Fachbereiche interessant sein könnten. GLDV-Mitgliedern und den Besucherinnen und Besuchern unserer Web-Seiten wäre ich jedenfalls für kurze Hinweise auf derartige Ereignisse dankbar.

Das Forum

Das zentrale Publikationsorgan der GLDV halten Sie gerade in der Hand, und die Web-Site trachtet hier keinesfalls nach Konkurrenz. Die elektronisch abrufbaren Seiten sind in der GLDV-Domäne primär als schnelle und auch vergängliche Informationsträger konzipiert. Nichtsdestoweniger bieten die Web-Seiten Einblick in das LDV-Forum. Zum einen geschieht dies dadurch, dass der Herausgeber einen stets aktuellen Blick in die Werkstatt auf seinem Web-Server in Darmstadt gestattet, in der auch Vorabveröffentlichungen von Beiträgen noch nicht erschienener Heft zu finden sind. Zum anderen sind beginnend mit Heft 2/1997 die PDF-Dateien der Druckvorlage von unserem Server abzurufen und können elektronisch

durchsucht werden. Das für das Öffnen von PDF-Dateien im Web-Browser erforderliche Plug-In ist von Adobe kostenlos erhältlich. Ein Link auf den Adobe-Web-Server ist bei uns eingerichtet.

Informationen von Absolvent-inn-en für Student-inn-en

Einige neue Seiten werden sich in Kürze in der GLDV-Web-Site finden: Eine Umfrage unter gerade noch Studierenden und frisch Examinierten in Fächern, die sich mit Computerlinguistik (, Linguistischer Datenverarbeitung, Sprachverarbeitung) oder auch allgemeiner mit Linguistik und anderen Nachbardisziplinen befassen, ist geplant. Natürlich auf freiwilliger Basis und auf Wunsch anonymisiert wollen wir als Entscheidungshilfe für Studierende und noch nicht Studierende die Absolventinnen und Absolventen nach ihren Fächerkombinationen, Abschlussarten, Studienorten, Semesterzahlen, Abschlussarbeiten und ihren Berufswünschen oder sich schon abzeichnenden Perspektiven fragen. Vielleicht kehrt ja die eine oder der andere auch später im Berufsleben nochmals zu dieser Seite zurück und kann etwas über den tatsächlichen Werdegang berichten. Für Interessentinnen und Interessenten sollen diese Informationen in aufbereiteter Form abrufbar sein. Die Ergebnisse dieser Umfrage ergänzen sinnvoll die Angaben, die der AK *Ausbildung und Berufsperspektiven* der GLDV in dem von Dr. Christian Wolff betreuten elektronischen Studienführer zusammenstellt und runden das Bild aus studentischer Sicht ab.

Die schwarzen Bretter der GLDV

Einen Versuch wert ist es, die Diskussion und Kommunikation in der GLDV elektronisch anzuregen, und zwar die außerhalb von Tagungen, Arbeitstreffen, Publikationen usw. Die bestehenden e-mail-Verteiler sind hier nicht das geeignete Mittel, da sie wahllos alle eingetragenen Mitglieder erreichen, ob sie nun am Diskussionsgegenstand interessiert sind und an der speziellen Diskussion teilnehmen wollen oder nicht. Abhilfe könnten abonnierbare Mailing-Listen darstellen oder Newsgroups bzw. deren WWW-Äquivalente. Wir haben uns für das letztgenannte in Form schwarzer Bretter entschieden. Als Eigenentwicklung[1] können sie von uns den Bedürfnissen, die sich hoffentlich noch zeigen werden, und zukünftigen Design-Entscheidungen leicht angepasst werden. Sie können mit geringem administrativen Aufwand eingerichtet werden, und zu ihrer Benutzung bedarf es nur dessen, was ohnehin jede Besucherin und jeder Besucher unserer WWW-Seiten hat: eines WWW-Browsers. Zwei Varianten solcher schwarzer Bretter bieten wir an.

Auf der ersten einfacheren Variante können primär kurze Nachrichten hinterlassen werden, die voneinander unabhängig sind. Das „Brett“ (das nicht schwarz ist) besteht aus zwei Frames. Im linken steht eine Liste der angehefteten Nachrichten mit Betreff und Name der Verfasserin oder des Verfassers, Datum und Uhrzeit der Nachricht. Der rechte Frame zeigt die Nachrichten selbst. Hier kann man sich mithilfe des Rollbalkens durch die Nachrichten bewegen oder durch Anklicken im linken Frame Nachrichten anspringen. Das Anheften einer neuen Nachricht geschieht selbsterklärend durch Auswahl des Links „[neuer Aushang]“. Gibt man hier auch eine e-mail-Adresse an, so eröffnet man den Leserinnen und Lesern die Möglichkeit einer direkten Antwort auf dem Wege der elektronischen Post.

Die zweite Brett-Variante, als Diskussions*forum* konzipiert, bietet die Möglichkeit thematischer Gliederung. Man kann neue Diskussionen eröffnen und in bestehende eingreifen. Hier steigt man über eine Liste von gerade diskutierten Themen ins Forum ein. Über das Link „[neues Thema]“ kann man jederzeit selbst eine neue Diskussion ins Leben rufen. Auf der Ebene der einzelnen Themen dann – nach Auswahl eines vorhandenen Themas oder Initiierung eines neuen – ist diese Art von schwarzen Brettern ähnlich strukturiert wie die erste. Der Titel, der einem neu eröffneten Thema gegeben wird, ist übrigens gleichzeitig auch der Betreff des ersten Beitrags.

Um die tiefe Verschachtelung von „Re:“s zu vermeiden, haben wir uns für eine rein chronologische Anordnung der Beiträge entschieden; nicht selten ist ja ein Beitrag ohnehin als Antwort auf mehrere vorangehende zu verstehen. Die „Dialogstruktur“ wird aber in der chronologischen Anordnung durch wechselseitige Links zwischen Beiträgen und den sie betreffenden Antworten offengelegt. Wir hoffen, dass sich auch der eine oder andere Arbeitskreis für die erste oder zweite Brettvariante entscheidet. (Nun, was das Design angeht, so handelt es sich natürlich um einen Metaphernbruch, wie man leicht beim Besuch der Seite sieht. Aber vielleicht ist das ja in diesem Fall erlaubt, wenn auch *schwarze Bretter* nach einer Orthographiereform nicht mehr schwarz sein müssen.)

Man „trifft“ sich in Zukunft hoffentlich häufiger in www.gldv.org !

Anmerkung

[1]Das Server-Programm stammt vom zweiten Autor dieses Beitrags, das „Äußere“ wurde von Frau Julia Nickel betreut.

Arbeitskreis-Profil „Quantitative Linguistik“

Der Arbeitskreis „Quantitative Linguistik“ besteht seit März 1991. Seine Aufgaben sind:

- Austausch aller Arten von Informationen über Aktivitäten im Forschungs- und Anwendungsbereich,
- Information über den Austausch von und Zugriffsmöglichkeiten auf Software und Daten und die Entwicklung neuer Methoden und Modelle
- Förderung der Kommunikation zwischen Forschung und Anwendung.

Der AK unterhält einen Rundbrief und trifft sich in unregelmäßigen Abständen. Die letzten größeren Treffen: Trierer Kolloquium zur quantitativen Linguistik (19.–20.10.'96) und Beteiligung an der QUALICO (Quantitative Linguistics Conference) in Helsinki im August 1997. Nächstes geplantes Treffen: Herbst '98 in Trier.

Neuer Arbeitskreis „Texttechnologie“

Der Vorstand der GLDV hat vor kurzem der Gründung des neuen Arbeitskreises *‚Texttechnologie‘* zugestimmt. Als Initiator möchte ich Sie über den Gegenstandsbereich dieses Arbeitskreises informieren und Sie zur Mitarbeit animieren!

In den achtziger Jahren ist mit der Standard Generalized Markup Language (SGML) eine Basis für die medienunabhängige Beschreibung von Textstrukturen und Annotationsssystemen entstanden, die in den letzten Jahren zu einer Vielzahl von Anwendungen – HTML ist darunter wohl die bekannteste –, Software-Systemen und abgeleiteten Standards geführt hat. Obwohl aber eine der Wurzeln von SGML in der Linguistik zu finden ist, sind zum Gebiet der maschinellen Sprach- und Textverarbeitung bisher kaum Verbindungen geschaffen worden. Der neu gegründete Arbeitskreis *Texttechnologie* hat sich zum Ziel gesetzt, die Kopplung von SGML-basierter Informationsverarbeitung, Linguistik und Sprachverarbeitung voranzutreiben, um damit die Entwicklung innovativer Textmodelle und inhaltsorientierter Textverarbeitung und -nutzung zu ermöglichen. Liora Alschuler bringt in ihrem einflußreichen Buch diese Verbindung gleich zu Beginn auf den Punkt:

> *If there is one single aspect that characterizes SGML […] it is that it puts the computing power of information technology behind the all-encompassing descriptive power of human language.*
>
> [Liora Alschuler, „ABCD ... SGML“. 1995, 1]

Im Fahrwasser von SGML sind eine Reihe weiterer Standards entstanden, die für diese Zielsetzung ebenfalls von Bedeutung sind: Die *Document Style Semantics and Specification Language* (DSSSL) erlaubt es, die Überführung von SGML-Instanzen in beliebige Präsentationsformate einschliesslich anderer SGML-Zielformate zu definieren. Die *Hypermedia/Time-based Structuring Language* (Hy-Time) stellt eine Konvention dar, wie Verweise in und zwischen Texten sowie zeitliche Abläufe und Synchronisationen in SGML-Instanzen auszudrücken sind. Für die Nutzung von SGML, DSSSL und HyTime im World-Wide Web sind darüber hinaus vereinfachte Versionen entwickelt worden oder gerade in der Entstehung: die *Extensible Markup Language* (XML), eine Vereinfachung von SGML, die *Extensible Linking Language* (XLL), eine Teilmenge von HyTime, sowie die *Extensible Style Language* (XSL), eine starke Vereinfachung von DSSSL.

Seit einiger Zeit bauen einige Kollegen aus dem Bereich Computerlinguistik und ich hier in Bielefeld den Bereich *‚Document Engineering/Texttechnologie‘* auf. Obwohl es in Deutschland inzwischen einige Gruppen mit ähnlichen Interes-

sen gibt, ist dieses Gebiet wie auch die gesamte SGML-basierte Texttechnologie im internationalen Vergleich in Deutschland noch recht schwach vertreten ist (s. z. B. die Zeitschrift „Text Technology“ oder Vorträge auf der GML/XML-97). Bei verschiedenen Gelegenheiten konnten wir feststellen, daß es bislang auch kein Forum gibt, in dem Kollegen mit entsprechenden Interessen virtuell oder real zusammenkommen können. Der Arbeitskreis Texttechnologie hat sich deshalb zum Ziel gesetzt, die Kommunikation unter den Wissenschaftlerinnen und Wissenschaftlern im Bereich der Texttechnologie durch eine Mailing-Liste und regelmäßige Workshops zu fördern, um dadurch auch nach außen größere Signifikanz zu erzielen.

Ich würde mich freuen, wenn Sie sich zu einer Mitarbeit in unserem Arbeitskreis entscheiden wuerden. Bitte schicken Sie mir in diesem Falle eine kurze Nachricht (lobin@lili.uni-bielefeld.de). Aktuelle Informationen zum Arbeitskreis sind in die GLDV-Seiten eingebunden (http://cl1.ikp.uni-bonn.de/GLDV/AKs/) oder unter dem URL http://coli.lili.uni-bielefeld.de/GLDV/AK-TT.html direkt zu erhalten.

Dr. Henning Lobin
WWW: http://coli.lili.uni-bielefeld.de/~lobin
Forschungschungsbereich ‚Document Engineering‘
Fakultät für Linguistik und Literaturwissenschaft
Universität Bielefeld, Postfach 10 01 31, 33501 Bielefeld
Tel. (0521) 106-3679, Fax (0521) 106-2996