Ontologie-basierte Hypertextsorten-Klassifikation

Georg Rehm

1 Einleitung

Der breite Einsatz computerlinguistischer Verfahren in Projekten zur Bewältigung der häufig zitierten Informationsflut beschränkt sich bislang meist auf Vorverarbeitungsprozesse (z.B. Wortstammreduktion oder Wortartenannotation zur Verbesserung von Information Retrieval-Algorithmen) oder klassische Anwendungen wie das automatische Textzusammenfassen oder die maschinelle Klassifikation eines Webdokuments in ein thematisches Kategoriensystem.

Dieser Beitrag stellt das Projekt Hypnotic (*Hypertexts and their Organisation into a Taxonomy by means of Intelligent Classification*) vor, in dem der Ansatz verfolgt wird, mit texttechnologischen und computerlinguistischen Verfahren HTML-Dokumente abstrakten *Hypertextsorten* zuzuordnen. In einem zweiten Schritt sollen – primär basierend auf dem Wissen, dass eine bestimmte Hypertextsorte vorliegt – generische Prozesse zur Informationsextraktion ausgeführt werden, um gezielt auf atomare und modulare Informationseinheiten, die in einem Dokument enthalten sind, zugreifen zu können. Zur Repräsentation und Manipulation der Informationen werden XML-basierte Formate und Standards eingesetzt.

Robuste Methoden zur automatischen Bestimmung von Hypertextsorten ermöglichten eine völlig neue Funktionalität für Suchmaschinen, indem der Benutzer die Möglichkeit erhielte, neben verschiedenen Stichwörtern auch die gewünschte(n) Hypertextsorte(n) der zu findenden Dokumente zu spezifizieren, beispielsweise „Texttechnologie“ und „XSLT“ in den Hypertextsorten *persönliche Homepage eines Wissenschaftlers* und *Wissenschaftlicher Artikel*. Auf diese Weise könnten Web-basierte Suchmaschinen unerwünschte Dokumente ausschließen, indem HTML-Dokumente mit abweichenden Hypertextsorten nicht in die Treffermenge aufgenommen werden. Wie bei vielen textlinguistisch untersuchten Textsorten (z.B. *Kochrezept*, *Memorandum* etc.) besitzen auch die Instanzen zahlreicher Hypertextsorten einen äußerst regulären Aufbau, so dass durch die Kenntnis der Hypertextsorte eines gegebenen Dokuments neuartige Analyse- und Explorationsmöglichkeiten entstehen.

Dieser Artikel stellt grundlegende Konzepte und Methoden des Projekts Hypnotic vor. Abschnitt 2 geht auf Aspekte der automatischen Erkennung ein, woraufhin Abschnitt 3 den modularen Aufbau von Hypertextsorten thematisiert. Abschließend werden exemplarisch die Ergebnisse einer Stichprobenanalyse dargestellt (Abschnitt 4). Ein zweiter Beitrag (Rehm, in diesem Band) diskutiert

vorwiegend Technologie-bezogene Aspekte des Projekts, stellt die Hypnotic-Korpusdatenbank vor, die ca. 4 Mio. deutschsprachige Dokumente der Webserver deutscher Hochschulen enthält, und geht auf den Aspekt der generischen Informationsextraktion ein.

2 Zur automatischen Erkennung von Hypertextsorten

Ein grundlegendes Ziel des Hypnotic-Systems soll es sein, in der Korpusdatenbank (vgl. Rehm 2001 und Rehm, in diesem Band) enthaltene HTML-Dokumente in eine hierarchisch angeordnete Ontologie von Hypertextsorten zu klassifizieren. Im Folgenden wird das Konzept der modular aufgebauten Hypertextsorten erläutert, an dessen Konstituenz unterschiedliche Ontologien beteiligt sind (Abschnitt 3). Abschnitt 4 stellt exemplarisch die Ergebnisse einer Stichprobenanalyse von 100 Dokumenten vor, die der Hypertextsorte *persönliche Homepage eines Wissenschaftlers* zugehörig sind. Zunächst wird jedoch eine Sammlung von Klassifikationsmerkmalen thematisiert, die in dem Hypnotic-System eingesetzt werden sollen und mit dem derzeitigen Stand der Kunst der maschinellen Klassifikation von Dokumenten in Text- bzw. Hypertextsorten kontrastiert.

Da die Hypertextsorten-Ontologie für die Domäne der Webserver deutscher Hochschulen bislang noch nicht abschließend spezifiziert wurde, betrachten wir die zugehörige Klassifikationskomponente derzeit als *black box*. In das Verfahren müssen jedoch neben dem in einem Dokument enthaltenen Text (bzw. Textfragmenten) auch strukturelle Merkmale einfließen. Dies zeigen vorläufige Resultate mit einem Modul, das auf manuell klassifizierten Trainingsdaten basiert und die Lernverfahren *Naive Bayes* sowie *k Nearest Neighbour* einsetzt. Für den Test wurden ca. 800 Dokumente als Trainingsgrundlage benutzt, die 80 flach angeordneten Hypertextsorten zugeordnet wurden. Etwa 500 unbekannte Dokumente wurden daraufhin automatisch klassifiziert, wobei die Präzision jedoch nur ca. 40% beträgt (bei einem Recall von etwa 60%). Durch Einbeziehung struktureller Eigenschaften sollte es möglich sein, diese Werte deutlich zu steigern. Hierzu wurden zahlreiche Merkmale gesammelt, deren konkrete Belegung für ein gegebenes Dokument in einer Weise extrahiert werden muss, die mit dem in Rehm (in diesem Band) dargestellten Ansatz zur Informationsextraktion vergleichbar ist und daher vermutlich auf den Ergebnissen der generischen Strukturanalyse basieren wird (vgl. hierzu Abb. 1 in Rehm, in diesem Band).

Der konkrete Einsatz der Klassifikationsmerkmale ist sehr stark von weiteren empirischen Analysen abhängig, weshalb derzeit noch keine Angaben über eine tatsächliche Implementierung sowie die Gewichtung einzelner Merkmale für die Klassifikationsaufgabe gemacht werden können. Da eine detaillierte Darstellung der Merkmale den Rahmen dieses Beitrags sprengte, sollen ledig-

lich Beispiele die grobe Vorgehensweise skizzieren (siehe hierzu auch Rehm 2002b):

Metadaten – URL eines Dokuments (beispielsweise ein ~-Zeichen, verschiedene „sprechende" Datei- oder Personennamen, vgl. Heißing 2000), HTTP Header (u.a. Last-Modified und Set-cookie), Dokumentgröße (in Bytes), Titel, Inhalt von `<meta>` Elementen etc.

HTML-Struktur – Globale Struktur des HTML-Elementbaums, bezieht u.a. mit ein:

- *Hyperlinks in einem Dokument bzw. einer Gruppe von Dokumenten* – Anzahl der Links, interne (zur gleichen Seite, zur Dokumentgruppe, zum gleichen Server, zu einem anderen Server innerhalb der gleichen Organisation) vs. externe Hyperlinks, Hypertextstruktur, Methode (HTTP, HTTPS, FTP etc.), Dateityp des Linkziels, Hypertextsorte des Ziels, Dateiname, Hyperlinkbezeichnung, Funktion (Haas/Grams 1998) und Position eines Links etc.
- *Inline-Graphiken* – Abmessungen von Graphiken (vgl. Rehm, in diesem Band), Datei- und Verzeichnisname, Inhalt, alternative Beschreibung, Format, Anzahl von Graphiken etc.
- *Interaktive Elemente* – HTML Formulare, JavaScript, Plug-Ins, Java Applets etc.

Linguistische Merkmale – Part-of-Speech-Frequenzen, Anzahl der Wörter und Sätze, Interpunktion, spezielle sprachliche Ausdrücke und Schlüsselwörter (an u.U. Hypertextsorten-spezifischen Positionen; Beispiele befinden sich u.a. in Rehm 2003, Roussinov *et al.* 2001, Toms/Campbell 1999, Haas/Grams 2000, de Saint-Georges 1998), Einordnung in ein Kontinuum, dessen Pole konzeptionelle Schriftlichkeit und konzeptionelle Mündlichkeit darstellen (hierzu Rehm 2003, Haase *et al.* 1997, Koch/Oesterreicher 1994).

Dokumentübergreifende Merkmale – Rekurrenz einzelner Hypertextsortenmodule (hierzu gehören z.B. Logos, Kopf- oder Fußzeilen), Position eines Dokuments bzgl. der Hypertextstruktur einer Dokumentgruppe (Wurzelknoten, Blattknoten) etc.

Dass eine maschinelle Textsorten-Klassifikation prinzipiell zufriedenstellende Ergebnisse liefert, zeigen u.a. Karlgren/Cutting (1994) und Kessler *et al.* (1997): Mit statistischen Verfahren werden Texte aus dem Brown Corpus in zwei bis vier Kategorien (z.B. *press*, *non-fiction*, *fiction*, *misc.* bei Karlgren/Cutting 1994) klassifiziert. Kessler *et al.* (1997) geben für ein sehr ähnliches Verfahren eine Präzision von 90% an. Bretan *et al.* (1998) benutzen zahlreiche

statistische Merkmale, die vornehmlich lexikalischer Natur sind, um einen C4.5-basierten Klassifikator (Quinlan 1993) zu trainieren, der Webseiten in die Kategorien *informal/private*, *public/commercial*, *journalistic materials*, *reports*, *other texts*, *interactive pages*, *discussions*, *link collections*, *FAQs* und *other listings and tables* einsortieren soll. Das Anwendungsszenario ist dabei die Hypertextsorten-getriebene Visualisierung der Ergebnisse einer Suchmaschine. Matsuda/Fukushima (1999) analysieren strukturelle Charakteristika von Webdokumenten, um Suchaufgaben im Kontext des Problemlösens zu unterstützen, wobei zur Berechnung eines „Dokumenttyps" (erkannt werden *product catalogue*, *online shop*, *advertisement*, *call for paper*, *links*, *FAQ*, *glossary*, *bulletin board* und *home page*) gewichtete, deskriptive Regeln eingesetzt werden. Die Autoren geben an, dass die Suche nach Dokumenttypen in konkreten Problemlösungsszenarien eine durchschnittliche Präzision von 88,9% aufweist, wohingegen die schlichte Suche nach Schlüsselwörtern lediglich 31,2% ergibt. Asirvatham/Ravi (2001) klassifizieren Dokumente in die Kategorien *information page*, *research page* und *personal home page*, wobei sowohl strukturelle als auch visuelle Merkmale, die aus eingebetteten Bildern gewonnen werden, in die Berechnung einfließen, die als Vergleich von Matrizen realisiert ist; die Autoren geben eine Präzision von etwa 87,8% an. Rauber/Müller-Kögler (2001) setzen eine automatische Genre-Analyse zur Visualisierung des Inhalts digitaler Bibliotheken ein. Hierbei wird eine Treffermenge als ein Bücherregal dargestellt, wobei das Aussehen eines Buches (Farbe, Form, Position im Regal, Stärke der Staubschicht etc.) u.a. das korrespondierende Genre und den letzten Zugriff repräsentiert; die Klassifikation findet mit Hilfe selbstorganisierender Karten statt, deren Eingabedaten aus Merkmalen bestehen, die auf die strukturelle und formatbezogene Heterogenität der Inhalte digitaler Bibliotheken zurückzuführen sind. Stamatatos *et al.* (2001) führen mit Hilfe von Worthäufigkeiten sowie Interpunktion eine Klassifizierung von griechischen Webdokumenten bzgl. Textsorte (u.a. *press editorial*, *reportage*, *academic prose*, *literature* und *recipes*) und Autorschaft durch. Eine Prämisse des Verfahrens ist, dass einzelne Kategorien in stilistischer Form homogen sein müssen. Insgesamt 22 Merkmale, die den jeweiligen Stil eines Textes markieren, werden aus den Ausgaben eines Werkzeugs berechnet, das u.a. Satzgrenzen bestimmt und ein partielles Parsing vornimmt. Die Klassifikation erfolgt primär mit Hilfe statistischer Methoden (multiple Regression sowie Diskriminanzanalyse). Als durchschnittliche Fehlerrate bei Tests mit jeweils 25 Dokumenten für 10 Genres geben die Autoren für beide Verfahren 0,18 an. Die Experimente von Finn *et al.* (2002) bzgl. der Aufgabe, Nachrichtenartikeln die Eigenschaften *Reportage* bzw. *Kommentar* mit Hilfe von Entscheidungsbäumen zuzuweisen, deuten darauf hin, dass eine Part-of-Speech-Repräsentation der Dokumente präzisere Resultate erbringt als ein purer „bag of words" Information-Retrieval-Ansatz oder eine Sammlung verschiedener linguistischer Eigenschaften eines Textes.

Bei der Evaluation mit unterschiedlichen Domänen (Fußball, Politik, Wirtschaft) schwankt die Präzision der eingesetzten Repräsentationen zwischen 60 und 90%. Bae-Lee/Myaeng (2002) benutzen Inhalts-spezifische Merkmale und ihre Gewichtungen für erweiterte tf idf-Statistiken, die mit Genre-spezifischen Merkmalen und Gewichtungen komplementiert werden. Auf diese Weise können in 533 persönlichen Homepages (dies stellt zugleich eines der betrachteten Genres dar), die in die inhaltlichen Kategorien *student*, *teacher/professor*, *company/employee* und *celebrity* partitioniert wurden, sowohl Genre- als auch Inhalts-spezifische Wörter aufgedeckt werden, wodurch eine Präzision von bis zu 90% erreicht wird. Dewdney *et al.* (2001) komplementieren ebenfalls eine thematische Analyse mit der Detektion von Genres (*ads*, *bulletin board*, *F.A.Q.*, *message board*, *radio*, *retuers*, *TV*), die auf 89 verschiedenen Merkmalen basieren, wobei *Support Vector Machines* (SVM), *Naive Bayes* und *C4.5* als Klassifikationsverfahren eingesetzt werden. Die Kombination der Eigenschaften „Thema“ und „Genre“ erreicht mit dem SVM-Klassifikator eine Präzision von 92%.

Die genannten Arbeiten zeigen zwar eine prinzipielle Realisierbarkeit der Klassifizierung von Dokumenten bzgl. ihrer Text- bzw. Hypertextsorte, jedoch müssen zahlreiche Fragestellungen genauer untersucht werden, beispielsweise die Skalierung dieser Verfahren, die lediglich von zwei bis zehn unterschiedlichen Sorten ausgehen, auf Hypertextsorten-Schemata, die mehrere Dutzend, evtl. sogar hierarchisch strukturierte Kategorien enthalten. Aus (text)linguistischer Sicht ist die Frage der Definition einer Hypertextsorte sowie ihre Abgrenzung von anderen Sorten von immenser Bedeutung, da hierdurch unmittelbar die Menge der Merkmale sowie ihr gewichteter Einfluss auf die Klassifikationsaufgabe computerlinguistisch motiviert wird; meines Wissens gehen Überlegungen dieser Art bislang nicht in existierende Prototypen ein, weshalb derartige theoretische Fragestellungen im Projekt Hypnotic bewusst im Zentrum stehen.[1]

1 Die automatische Klassifikation von Webseiten in Hypertextsorten wird m.W. derzeit von keinem öffentlich zugänglichen System vorgenommen oder unterstützt, und auch im Forschungsbereich des „Web Searching“ wird die Idee bislang nicht thematisiert, vgl. etwa Jansen/Pooch (2001).

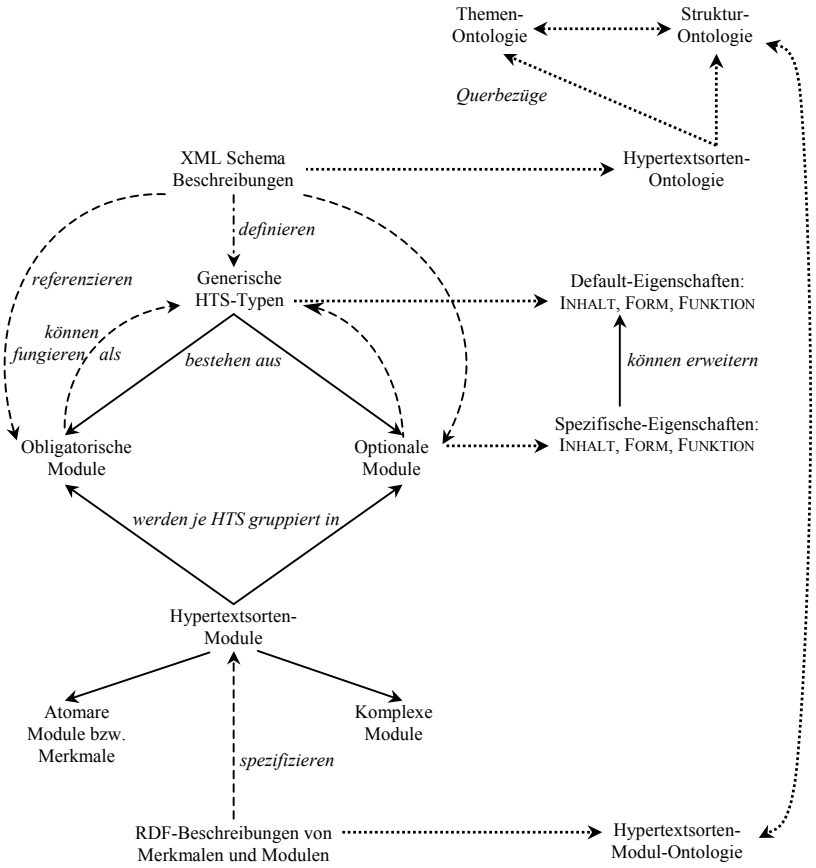

Abbildung 1: Aufbau und Repräsentation generischer Hypertextsorten-Typen durch Hypertextsorten-Module und assoziierte Merkmale (vgl. Tab. 1)

3 Hypertextsorten und Hypertextsorten-Module

Hypertextsorten sind konventionalisierte textuelle Strukturen, die sich seit der Etablierung des World Wide Webs Anfang der neunziger Jahre kontinuierlich gebildet haben bzw. noch immer bilden. Sie sind konzeptionell unmittelbar mit Textsorten vergleichbar, d.h. innerhalb des Webs formieren sich zahlreiche metatextuelle Schemata, die bestimmten kommunikativen Situationen zuzuordnen bzw. in diese eingebettet sind und über jeweils charakteristische Eigen-

schaften bzgl. der Merkmale Inhalt, Form und Funktion verfügen.[2] Die Terminologie in diesem bislang kaum erforschten Thema ist noch uneinheitlich; in der amerikanischen Forschung ist häufig von *digital genres* oder auch *web genres* die Rede, wohingegen sich im Deutschen der Begriff *Hypertextsorte* aufgrund der konzeptionellen Nähe zu dem hierzulande umfangreich erforschten Gebiet der Text- und Textsortenlinguistik anbietet. Auf die Fragen nach der Anzahl von Hypertextsorten, ihren konstituierenden und distinktiven Merkmalen, evtl. existierenden, unterschiedlichen Typologien etc. liegen bislang nur wenige Antworten vor (vgl. etwa Brandl 2002). Beispielhaft können verschiedene Studien genannt werden, die zufällige Stichproben zwischen 100 und 1000 HTML-Dokumenten von regulären Suchmaschinen beziehen und versuchen, diese Dokumente anhand intuitiv festgelegter Eigenschaften in unterschiedliche Hypertextsorten einzuteilen (vgl. u.a. Shepherd/Watters 1999, Dillon/Gushrowski, 2000, Crowston/Williams 2000, Haas/Grams 2000, Roussinov *et al.* 2001, Rehm 2002b enthält eine Übersicht).[3]

In den bislang vorgelegten Arbeiten (s.o.) befinden sich fast immer Hinweise darauf, dass sich eine Klassifizierung von HTML-Dokumenten in konkrete Hypertextsorten aufgrund der zahlreichen konstituierenden Elemente sehr schwierig gestaltet, weshalb häufig davon ausgegangen wird, dass es sich bei HTML-Dokumenten nicht um monolithische, sondern um hochgradig modulare Einheiten handelt (Rehm 2002b, Haas/Grams 2000), die auf primitiven, generischen Hypertextsorten-Typen basieren, die wiederum den grundlegenden Aufbau einer Hypertextsorte abstrakt charakterisieren (vgl. Abb. 1). Hypertextsorten-Typen spezifizieren zwei Sorten von Hypertextsorten-Modulen: Obligatorische Module müssen in einer Hypertextsorten-Instanz enthalten sein, damit ein Dokument zu der entsprechenden Hypertextsorte gehört; optionale Module können zusätzlich vorhanden sein, um Varietäten zu erlauben. Hypertextsorten-

2 Bates/Lu (1997) untersuchen 114 persönliche Homepages und kommen in einer Zeit, in der sich Hypertextsorten noch in einer frühen Formierungsphase befanden, zu dem Schluss: „[...] it appears that the form and content of personal home pages on the World Wide Web is still quite open and various. [T]he public social form known as the 'personal home page' has not yet developed a fully standardised character and social role, recognised by all. The home page as a social institution is still very much under development."

3 Mir sind keine Arbeiten bekannt, die sich in der Tradition der deutschen Textlinguistik mit Hypertextsorten beschäftigen. Betrachtet man aktuelle Beiträge zur Textlinguistik, z.B. Heinemann (2000), fällt auf, dass hier im Kontext der digitalen Informations- und Kommunikationsdienste häufig das Medium, genauer gesagt, der Service selbst, als „Textsorte" fehlinterpretiert wird: „[...] im Umgang mit neuen Textsorten (E-Mail, Hypertext)" (S. 507), wohingegen bereits in zahlreichen Arbeiten zur computervermittelten Kommunikation ein sehr breites Spektrum von Varietäten nachgewiesen wurde, die z.B. innerhalb der elektronischen Post von den Benutzern intuitiv eingesetzt werden, vgl. u.a. Runkehl *et al.* (1998, S. 51). Es liegt nahe, derartige Varietäten auch für den Bereich des Hypertext-Dienstes World Wide Web anzunehmen (speziell zum Verhältnis von E-Mail zu persönlichen Homepages siehe Rehm 2002a).

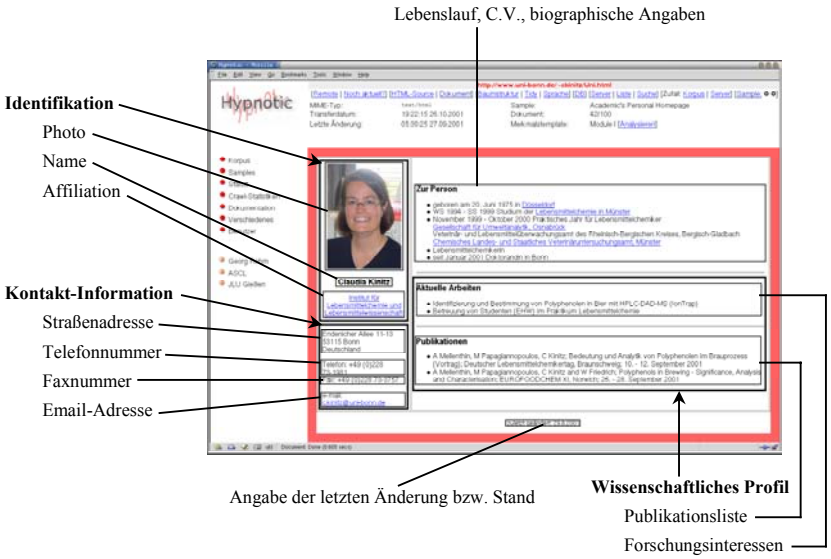

Abbildung 2: Komplexe Hypertextsorten-Module (in Fettdruck) und ihre konstituierenden atomaren Merkmale bzw. Module am Beispiel einer Instanz der Hypertextsorte persönliche Homepage eines Wissenschaftlers (vgl. Tab. 1), die in der Hypnotic-Korpusdatenbank enthalten ist

Typen besitzen eindeutig kennzeichnende, prototypische Eigenschaften bzgl. der Merkmale Inhalt, Form und Funktion.[4] Konkret spezifiziert stellen diese Eigenschaften die typische Definition einer Hypertextsorte dar, die durch das Vorhandensein optionaler Module entsprechend modifiziert werden kann (vgl. Rehm 2002b, für eine konzise Definition der Hypertextsorte persönliche Homepage eines Wissenschaftlers; ein Querschnittsüberblick zum Thema „persönliche Homepages" befindet sich in Döring 2001). Da die obligatorischen Hypertextsorten-Module fester Bestandteil einer Hypertextsortendefinition sind, besitzen generalisierte Gruppen von optionalen Modulen einen derartigen speziellen Status, um Varietäten auch inhaltlich, formal und funktional erfassen zu können.[5] Hypertextsorten-Module können komplex oder atomar sein, wobei komplexe Module aus atomaren Modulen bzw. Merkmalen zusammengesetzt

4 Diese drei Merkmale wurden initial gewählt, da sie die grundlegenden Möglichkeiten der Spezifizierung unterschiedlicher Text- und somit auch Hypertextsorten darstellen.

5 Ein optionales Hypertextsorten-Modul *Suchformular* könnte in einer Hypertextsorten-Instanz vorkommen und somit durch den definierten Status einer hohen Interaktivität vornehmlich das Merkmal „Funktion" modifizieren.

sind, wie beispielsweise *Wissenschaftliches Profil* (vgl. Cronin *et al.* 1998), das sich innerhalb des generischen Hypertextsorten-Typs *persönliche Homepage eines Wissenschaftlers* u.a. aus den atomaren Modulen *Publikationsliste* und *Forschungsinteressen* konstituiert, vgl. Abbildung 2 für ein Beispiel. Hypertextsorten-Module können darüber hinaus auch als eigenständige Hypertextsorten-Typen fungieren, etwa die *Publikationsliste*, die sowohl – *eingebettet* als Modul – in einer Hypertextsorten-Instanz existent sein kann oder – *verlinkt* als Modul – in einem separaten physikalischen Dokument vorhanden sein kann; hierbei gehört die verlinkte Datei, wenn sie isoliert betrachtet wird, zum generischen Hypertextsorten-Typ *Publikationsliste*.

Zur Repräsentation der generischen Hypertextsorten-Typen sollen XML Schema Beschreibungen (Fallside et al. 2001) eingesetzt werden, sobald für die Untersuchungsdomäne alle identifizierten Typen analysiert und spezifiziert sind. Auf diese Weise können Vererbungsmechanismen die Reduplikation von Definitionen vermeiden, so dass die Hypertextsorte persönliche Homepage eines Wissenschaftlers als eine spezielle Varietät der generischeren Hypertextsorte persönliche Homepage repräsentierbar ist. Ein weiterer Vorteil dieses validierbaren Formats besteht in der Möglichkeit, maschinell annotierte Dokumente gegen die korrespondierende XML Schema Beschreibung zu parsen, um fehlerhafte Auszeichnungen unmittelbar aufzudecken (vgl. Rehm 1999). Auf der Erkennungsseite werden RDF-Beschreibungen benutzt, um die beteiligten Merkmale von Hypertextsorten-Modulen zu repräsentieren.[6] Der Merkmalsbegriff bezieht sich hier primär auf inhaltliche, formale und funktionale Eigenschaften, die maschinell detektierbar sind. Die somit entstehende Ontologie von Hypertextsorten-Modulen wird mit der Hypertextsorten-Ontologie verschaltet, um die abstrakte Definitions- mit der konkreten Erkennungsebene zu verknüpfen. Weitere Querbezüge werden vermutlich zu einer thematischen[7] und einer strukturellen Ontologie bestehen müssen: Die Struktur-Ontologie bezieht sich dabei auf die strukturelle Organisation von Hochschulen, die unmittelbar durch die Strukturierung der im World Wide Web verfügbaren Inhalte reflektiert wird, etwa eine Unterteilung in Fachbereiche bzw. Fakultäten, die wiederum aus Instituten und einem Dekanat bestehen. Abbildung 3 zeigt einen Ausschnitt der Struktur-Ontologie, die aus einer Analyse von 727 Dokumenten hervorgegangen ist. Diese Dokumente bestehen aus den Einstiegsseiten von 35 in der Korpusdatenbank enthaltenen Universitäten sowie allen im Korpus verfügbaren

6 Ein ähnlicher Ansatz wird im Kontext der Agenten-basierten Identifikation von Artikeln aus digital verfügbaren Tageszeitungen in dem Projekt VIPAR verfolgt, siehe Potok *et al.* (2002).

7 Als Basis der thematischen Ontologie könnte z.B. eines der Klassifikationssysteme UDC (Universal Decimal Classification, http://www.udcc.org) oder DDC (Dewey Decimal Classification, http://www.oclc.org/dewey/) dienen.

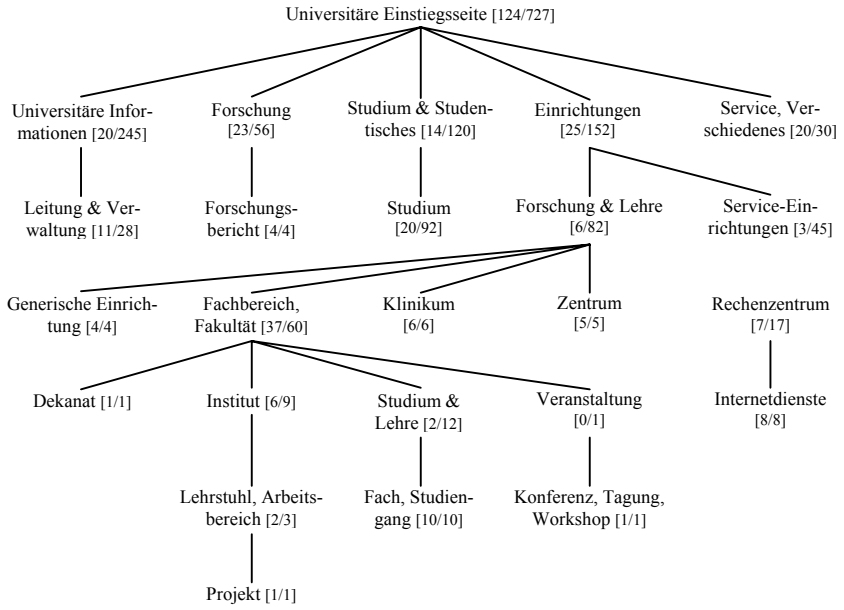

Abbildung 3: Ausschnitt der Struktur-Ontologie für die Domäne „Deutsche Hochschule" und Häufigkeiten (in einem Knoten/unterhalb eines Knotens) der jeweiligen Ebene in einer Stichprobe von 727 Dokumenten der ersten Verlinkungsebene von 35 Universitäten

Dokumenten, auf die von den Einstiegsseiten aus verwiesen wird. Die an den Knoten der Ontologie befindlichen Tupel stellen Frequenzangaben dar: Der erste Wert kennzeichnet die Menge der Dokumente innerhalb eines Knotens, der zweite Wert markiert alle Dokumente in und unterhalb eines Knotens.[8] Diese Art der Analyse, die semiautomatisch mit Hilfe der Korpusdatenbank durchgeführt wird, stellt zugleich die Methode dar, mit der unterschiedliche Typen von Hypertextsorten iterativ durch die Inspektion von Stichproben und die Aufstellung von Merkmalslisten identifiziert werden. Die Analyse einer initialen Stichprobe von 200 zufällig ausgewählten Dokumenten resultierte in 84 unterschiedlichen Hypertextsorten[9] bzw. Hypertextsorten-Varietäten, woraufhin die bereits angesprochene Analyse der Einstiegsseiten sowie der Dokumente, auf die diese Seiten verweisen, den Aufbau der oberen Ebenen der Hypertextsorten-Ontologie ergeben hat. Abschließend sollen 2000 zufällig

8 Der erste Wert des Tupels an der Wurzel des Baumes beträgt nicht 35, weil an dieser Stelle eine Strukturtypisierung vorgenommen wird, so dass sich nicht nur alle 35 universitären Einstiegsseiten innerhalb des Wurzelknotens befinden.

9 Ein Ausschnitt dieser initialen Ontologie ist in Rehm (2002b) dargestellt.

ausgewählte Dokumente tieferer Ebenen die empirisch aufgebaute Ontologie vervollständigen.

4 Analyse einer Stichprobe von Dokumenten

Die grundlegende Methodik zum Aufbau abstrakter Spezifikationen von Hypertextsorten stellt die Inspektion von Stichproben dar. Hierbei unterstützt die Hypnotic-Korpusdatenbank (s. Rehm, in diesem Band) in vielen Fällen sowohl die Phasen der Generierung als auch der Analyse mit Hilfe eines Webbrowsers (Rehm 2003). Im Folgenden soll die Definition der Hypertextsorte *persönliche Homepage eines Wissenschaftlers* genauer betrachtet werden. Die entsprechende Stichprobe basiert auf 100 Dokumenten, die semiautomatisch mit Hilfe des Hypnotic-Systems zusammengestellt wurden. Zunächst wurde eine nach Häufigkeiten sortierte Liste derjenigen Webserver generiert, die persönliche Homepages nach der Tilde-Konvention (z.B. http://www.uni-giessen.de/~g91063/) anbieten.[10] Auf diese Weise sind u.a. die Webserver auffindbar, deren Namen darauf hindeuten, dass ausschließlich Homepages von Wissenschaftlern vorhanden sind (z.B. 763 Homepages auf dem Webserver staff-www.uni-marburg.de), um die Zusammenstellung der Stichprobe dadurch zu vereinfachen, dass studentische Seiten unmittelbar ausgeschlossen werden. Daraufhin wurden von den ersten fünf Webservern 100 Dokumente in das Sample aufgenommen, sofern verschiedene Bedingungen erfüllt sind: Das Dokument ist (a) Einstiegsseite der persönlichen Homepage eines Wissenschaftlers, (b) deutschsprachig, (c) gehört eindeutig einer Person (z.B. im Gegensatz zu einer Arbeitsgruppe), (d) hat primär mit der Tätigkeit des Autors an der Universität zu tun und (e) setzt keine Framesets ein.

10 Derartige Frequenzlisten sind auf dem Korpusdatenbank-Server mit Hilfe von SQL-Querys und UNIX Shell-Skripten, die auf der Ausgabe der relationalen Korpusdatenbank operieren, mit einfachen Mitteln erstellbar.

Tabelle 1: Spezifikation der Hypertextsorte persönliche Homepage eines Wissenschaftlers durch Angabe der obligatorischen und optionalen Module, die auf den Häufigkeitsangaben der Stichprobenanalyse (vgl. Abschnitt 4) basieren

Ebene	*Bezeichnung/Erläuterung*	*Status*	*Vorkommen in dieser Hypertextsorte*	*Häufigkeit in der Stichprobe*
Atomares Modul	**Explizite Begrüßung**	generell	optional	14
Kompl. Modul	**Identifikation**	generell	obligatorisch	–
Merkmal	*Name des Homepage-Besitzers*	generell	obligatorisch	100
Merkmal	*... begleitet von Titelangabe*	spezifisch	obligatorisch	69
Merkmal	*... begleitet von Tätigkeitsangabe*	generell	optional	27
Merkmal	*... begleitet von Affiliation*	generell	obligatorisch	34
Merkmal	*... begleitet von Photo des Autors*	generell	obligatorisch	54
Kompl. Modul	**Eigenständige Affiliation**	generell	obligatorisch	–
Merkmal	*Name der Universität im Klartext*	generell	obligatorisch	75
Merkmal	*Logo der Universität*	generell	optional	16
Atomares Modul	**Alternative Version (andere Sprache)**	generell	optional	75
Kompl. Modul	**Kontakt-Informationen**	generell	obligatorisch	–
Merkmal	*Straßenadresse (Universität, Straße, PLZ, Stadt, Land etc.)*	generell	obligatorisch	90
Merkmal	*Explizite Postadresse*	generell	optional	8
Merkmal	*Telefonnummer*	generell	obligatorisch	86
Merkmal	*Telefonnummer (Sekretariat)*	generell	optional	7
Merkmal	*Faxnummer*	generell	obligatorisch	66
Merkmal	*Email-Adresse*	generell	obligatorisch	98
Merkmal	*Angabe der URL dieser Homepage*	generell	optional	4
Merkmal	*Zimmernummer*	generell	obligatorisch	30
Merkmal	*SMS senden*	generell	optional	1
Merkmal	*PGP Public Key bzw. Fingerprint*	generell	optional	2
Merkmal	*X.500 Eintrag*	generell	optional	2
Merkmal	*Informationen zur Anreise*	generell	optional	2
Merkmal	*Sprechstunden*	spezifisch	optional	25
Merkmal	*Adresse (privat)*	generell	optional	18
Merkmal	*Telefonnummer (privat)*	generell	optional	22
Merkmal	*Mobiltelefonnummer (privat)*	generell	optional	3
Merkmal	*Faxnummer (privat)*	generell	optional	7
Merkmal	*Email-Adresse (privat)*	generell	optional	5
Merkmal	*URL der privaten Homepage*	generell	optional	2
Kompl. Modul	**Kontakt-Informationen (Sekretariat)**	spezifisch	optional	–
Merkmal	*Name*	generell	obligatorisch	8
Merkmal	*Straßenadresse*	generell	optional	3
Merkmal	*Zimmernummer*	generell	optional	4
Merkmal	*Öffnungszeiten*	generell	optional	5
Merkmal	*Telefonnummer*	generell	optional	6
Merkmal	*Faxnummer*	generell	optional	6

Merkmal	*Email-Adresse*	generell	optional	6
Kompl. Modul	**Kontakt-Informationen (Mitarbeiter)**	spezifisch	optional	–
Merkmal	*Name*	generell	obligatorisch	7
Merkmal	*Auflistung mehrere Einträge*	meta	optional	6
Merkmal	*Adresse*	generell	optional	2
Merkmal	*Zimmernummer*	generell	optional	3
Merkmal	*Telefonnummer*	generell	optional	4
Merkmal	*Email-Adresse*	generell	optional	4
Merkmal	*Begleitet von Namen von Hilfskräften*	generell	optional	2
Kompl. Modul	**Universitäres Profil**	spezifisch	obligatorisch	–
Merkmal	*Angaben zu Lehrveranstaltungen*	spezifisch	obligatorisch	49
Merkmal	*Funktionen innerhalb der Universität (z.B. Gremienarbeit)*	spezifisch	optional	7
Merkmal	*Allgemeine Studienhinweise*	spezifisch	optional	3
Merkmal	*Angebotene Abschlussarbeiten*	spezifisch	optional	2
Kompl. Modul	**Wissenschaftliches Profil**	spezifisch	obligatorisch	–
Merkmal	*Publikationsliste*	spezifisch	obligatorisch	71
Merkmal	*Forschungsinteresseren*	spezifisch	obligatorisch	50
Merkmal	*Forschungsprojekte*	spezifisch	optional	22
Merkmal	*Prominent platzierte Bücher bzw. Zeitschriften*	spezifisch	optional	6
Merkmal	*Liste von Vorträgen bzw. Präsentationen*	spezifisch	optional	5
Merkmal	*Mitgliedschaft in Fachverbänden*	spezifisch	optional	4
Merkmal	*Technologietransfer*	spezifisch	optional	1
Atomares Modul	**Lebenslauf, C.V., biographische Angaben**	generell	obligatorisch	60
Atomares Modul	**Interessante Links**	generell	optional	12
Kompl. Modul	**Relevante Links**	generell	optional	–
Merkmal	*Link zum eigenen Institut bzw. Arbeitsbereich*	spezifisch	obligatorisch	49
Merkmal	*Link zur Einstiegsseite der eigenen Universität*	spezifisch	obligatorisch	36
Merkmal	*Link zur eigenen Fakultät bzw. Fachbereich*	spezifisch	optional	23
Atomares Modul	**Angabe der letzten Änderung bzw. Stand**	universal	obligatorisch	42
Atomares Modul	**Counter, Zugriffszähler**	universal	optional	11
Atomares Modul	**Gästebuch**	universal	optional	1

Tabelle 1 stellt einen Ausschnitt des Resultats der Stichprobenanalyse dar, wobei die ersten Hypertextsorten-Module im Folgenden erläutert werden sollen. Der generische Hypertextsorten-Typ *persönliche Homepage eines Wissenschaftlers* besteht aus atomaren und komplexen Modulen, wobei die Reihenfolge der Darstellung in der Tabelle eine Approximation der „kanonischen" Sequenz in den Dokumenten der Stichprobe ist. Nach einer optionalen *expliziten Begrüßung*, die generell in allen Hypertextsorten der Gruppe *persönliche Homepage* enthalten sein darf, folgt das ebenfalls generelle Modul *Identifikation*, das aus einzelnen Merkmalen besteht, von denen der *Name des Homepage-Besitzers* das zentrale Modul darstellt[11]; innerhalb dieses komplexen

11 Der Name des Homepage-Besitzers ist als einziges Modul in jedem einzelnen Dokument der Stichprobe enthalten; an zweiter Stelle folgt die Email-Adresse mit 98 Vorkommen (Bates/Lu 1997 berichten, dass in einem Sample persönlicher Homepages nur in 79% der Dokumente die

Moduls ist lediglich die Titelangabe spezifisch für die Hypertextsorte *persönliche Homepage eines Wissenschaftlers*. Die *eigenständige Affiliation* kennzeichnet die Organisation bzw. Institution, für die der Autor der Seite tätig ist, üblicherweise geschieht dies (75 Vorkommen) durch die Nennung der Hochschule im Klartext oder – bei 16 Vorkommen ersetzend bzw. ergänzend – durch die Einbindung des Logos oder Siegels der Universität. Die weiteren komplexen Module sind *Kontakt-Information*, das zahlreiche Möglichkeiten der Kontaktaufnahme subsumiert, *Kontakt-Information (Sekretariat)* sowie *(Mitarbeiter)* und die beiden zentralen Module *Universitäres Profil* und *Wissenschaftliches Profil*, die einerseits *Angaben über Lehrveranstaltungen*, *Funktionen innerhalb der Universität* und *Studienhinweise*, andererseits eine *Publikationsliste*, *Forschungsinteressen* und *Forschungsprojekte* enthalten. Die drei letzten atomaren Module (*Angabe der letzten Änderung*, *Zugriffszähler*, *Gästebuch*) haben einen universellen Status, d.h. sie können prinzipiell in *jeder* Hypertextsorte vorkommen, ihre Existenz ist nicht nur auf die Hypertextsorten-Gruppe *persönliche Homepage* beschränkt.

5 Zusammenfassung und Ausblick

Dieser Beitrag stellt grundlegende theoretische Konzepte vor, die derzeit in einem System zur automatischen Klassifizierung von Webdokumenten in Hypertextsorten implementiert werden. Von zentraler Bedeutung sind hierbei verschiedene miteinander vernetzte Ontologien, die – realisiert mit Hilfe von XML Schema – sowohl den modularen Aufbau abstrakter Hypertextsorten-Typen, unterschiedliche Klassifikationsmerkmale (in Form von RDF-Beschreibungen) als auch die Struktur der inhaltlichen Domäne (Webserver deutscher Hochschulen) betreffen. Die entsprechenden Ontologien werden iterativ mittels empirischer Methoden gefüllt, die insbesondere auf der Analyse von Stichproben basieren, die wiederum mit Hilfe der Web-Oberfläche der Hypnotic-Korpusdatenbank generiert und verwaltet werden.

Die zukünftige Arbeit bezieht sich vornehmlich auf die Analyse der in Abschnitt 3 angesprochene Stichprobe 2000 zufällig ausgewählter Dokumente und einer anschließenden Kopplung der Resultate mit den bereits untersuchten Stichproben, die sich insbesondere auf die obere Ebene der Hypertextsorten-

Namen der Autoren vorhanden waren, die Email-Adresse war in 92,1% der Dokumente angegeben). Hier wird auch deutlich, dass komplexe Module keine Frequenzangaben besitzen, da sie aus atomaren Modulen zusammengesetzt sind. Ein Merkmal bzw. Modul wird in dieser Analyse als obligatorisch eingestuft, sofern 30 oder mehr Vorkommen in der Stichprobe enthalten sind. Eine Ausnahme stellt das Modul alternative Version in einer anderen Sprache dar, denn durch das Fehlen einer alternativen Sprachversion einer Homepage erfährt die ursprünglich vorhandene Version keine unmittelbare Einschränkung bzgl. den Eigenschaften Inhalt, Form oder Funktion.

Ontologie sowie auf den Bereich der persönlichen Homepages beziehen. Weiterhin steht die Implementation des Prototypen im Vordergrund, der – neben der Klassifikation von Webdokumenten in Hypertextsorten – auch in der Lage sein soll, basierend auf dem Klassifikationsergebnis eine generische Informationsextraktion durchzuführen. Dieser Gesichtspunkt wird in Rehm (in diesem Band) in detaillierter Form dargestellt.

6 Literatur

Asirvatham, Arul Prakash und Kranthi Kumar Ravi: Web Page Classification Based on Document Structure. Technischer Bericht. International Institute of Information Technology, Hyderabad 2001. [http://www.iiit.net/stud_pub.htm]

Bae-Lee, Yong und Sung Hyon Myaeng: Text Genre Classification with Genre-Revealing and Subject-Revealing Features. In: Proceeding of the 25th Annual International Conference on Research and Development in Information Retrieval (SIGIR '02). Tampere: ACM Press 2002, S. 145-150.

Bates, Marcia J. und Shaojun Lu: An Exploratory Profile of Personal Home Pages: Content, Design, Metaphors. In: Online & CD ROM Review 21 (6), 1997, S. 331-340.

Brandl, Annette: Webangebote und ihre Klassifikation – Typische Merkmale aus Experten- und Rezipientenperspektive. (Angewandte Medienforschung – Schriftenreihe des Medien Instituts Ludwigshafen, Nummer 21). München: R. Fischer 2002.

Bretan, Ivan, Johan Dewe, Anders Hallberg, Niklas Wolkert und Jussi Karlgren: Web-Specific Genre Visualization. In: Proceedings of the 3rd WebNet Conference (WebNet '98), Orlando 1998.

Cronin, Blaise, Herbert W. Snyder, Howard Rosenbaum, Anna Martinson und Ewa Callahan: Invoked on the Web. In: Journal of the American Society for Information Science 49 (14), 1998, S. 1319-1328.

Crowston, Kevin und Marie Williams: Reproduced and Emergent Genres of Communication on the World Wide Web. In: The Information Society 16 (3), 2000, S. 201-215.

de Saint-Georges, Ingrid: Click Here if You Want to Know Who I Am. Deixis in Personal Homepages. In: Proceedings of the 31st Hawaii International Conference on Systems Sciences (HICSS-31), IEEE 1998.

Dewdney, Nigel, Carol van Ess-Dykema und Richard MacMillan: The Form is the Substance: Classification of Genres in Text. In: Proceedings of Workshop on Human Language Technology and Knowledge Management. Toulouse: Association for Computational Linguistics 2001.

Dillon, Andrew und Barbara A. Gushrowski: Genres and the Web: Is the Personal Home Page the First Uniquely Digital Genre? In: Journal of the American Society for Information Science 51 (2), 2000, S. 202-205.

Döring, Nicola: Persönliche Homepages im WWW. In: Medien & Kommunikationswissenschaft 49 (3), 2001, S. 325-349.

Fallside, David C., Henry S. Thompson, David Beech, Murray Maloney, Noah Mendelsohn, Paul V. Biron und Ashok Malhotra: XML Schema. Technische Spezifikation, World Wide Web Consortium. W3C Recommendation. 2001. Besteht aus Part 0 (Primer), Part 1 (Structures), Part 2 (Datatypes). [http://www.w3.org/XML/Schema]

Finn, Aidan, Nicholas Kushmerick und Barry Smyth: Genre Classification and Domain Transfer for Information Filtering. In: *Crestani, F., M. Girolami und C.J. van Rijsbergen* (Hg.): Advances in Information Retrieval – 24th BCS-IRSG European Colloquium on IR Research. (Lecture Notes in Computer Science, number 2291). Berlin u.a.: Springer 2002, S. 353-362.

Haas, Stephanie W. und Erika S. Grams: A Link Taxonomy for Web Pages. In: *Preston, C.* (Hg.): Proceedings of the 61st Annual Meeting of the American Society for Information Science 1998, S. 485-495.

Haas, Stephanie W. und Erika S. Grams: Readers, Authors, and Page Structure – A Discussion of Four Questions Arising from a Content Analysis of Web Pages. In: Journal of the American Society for Information Science 51 (2), 2000, S. 181-192.

Haase, Martin, Michael Huber, Alexander Krumeich und Georg Rehm: Internetkommunikation und Sprachwandel. In: *Weingarten, Rüdiger* (Hg.): Sprachwandel durch Computer. Opladen: Westdeutscher Verlag 1997, S. 1-85.

Heinemann, Wolfgang: Textsorte – Textmuster – Texttyp. In: *Brinker, Klaus, Gerd Antos, Wolfgang Heinemann und Sven F. Sager* (Hg.): Text- und Gesprächslinguistik. (Handbücher für Sprach- und Kommunikationswissenschaft (HSK), Bd. 16). Berlin, New York: de Gruyter 2000, S. 507-523.

Heißing, Christian: Klassifizieren von URLs durch Generalisierung von Pfadnamen. Magisterarbeit im Studiengang Computerlinguistik und Künstliche Intelligenz, Institut für Semantische Informationsverarbeitung, Universität Osnabrück 2000.

Jansen, Bernard J. und Udo Pooch: A Review of Web Searching Studies and a Framework for Future Research. In: Journal of the American Society for Information Science and Technology 52 (3), 2001, S. 235-246.

Karlgren, Jussi und Douglass Cutting: Recognizing Text Genres with Simple Metrics Using Discriminant Analysis. In: COLING 94 – The 15th International Conference on Computational Linguistics, Bd. 2. Kyoto, Japan: Association for Computational Linguistics 1994, S. 1071-1075.

Kessler, Brett, Geoffrey Nunberg und Hinrich Schütze: Automatic Detection of Text Genre. In: Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th Meeting of the European Chapter of the Association for Computational Linguistics. San Francisco: Morgan Kaufmann 1997, S. 32-38.

Koch, Peter und Wulf Oesterreicher: Schriftlichkeit und Sprache. In: *Günther, H. und O. Ludwig* (Hg.): Schrift und Schriftlichkeit. (Handbücher für Sprach- und Kommunikationswissenschaft (HSK), Bd.10). Berlin, New York: de Gruyter 1994, S. 587-604.

Matsuda, Katsushi und Toshikazu Fukushima: Task-Oriented World Wide Web Retrieval by Document Type Classification. In: Proceedings of the International Conference on Information and Knowledge Management (CIKM '99), Kansas City 1999, S. 109-113.

Potok, Thomas E., Mark T. Elmore, Joel W. Reed und Nagiza F. Samatova: An Ontology-based HTML to XML Conversion Using Intelligent Agents. In: Proceedings of the 35th Hawaii International Conference on System Sciences (HICSS-35), Big Island, Hawaii 2002.

Quinlan, J. Ross: C4.5: Programs for Machine Learning. San Mateo: Morgan Kaufmann 1993.

Rauber, Andreas und Alexander Müller-Kögler: Integrating Automatic Genre Analysis into Digital Libraries. In: *Fox, E. A. und C. L. Roanoke Borgman* (Hg.): Proceedings of the ACM/IEEE Joint Conference on Digital Libraries. ACM Press 2001, S. 1-10.

Rehm, Georg: Automatische Textannotation – Ein SGML- und DSSSL-basierter Ansatz zur angewandten Textlinguistik. In: *Lobin, Henning* (Hg.): Text im digitalen Medium – Linguistische Aspekte von Textdesign, Texttechnologie und Hypertext Engineering. Wiesbaden: Westdeutscher Verlag 1999, S. 179-195.

Rehm, Georg: `korpus.html` – Zur Sammlung, Datenbank-basierten Erfassung, Annotation und Auswertung von HTML-Dokumenten. In: *Lobin, Henning* (Hg.): Proceedings der GLDV Frühjahrstagung 2001. Giessen: Gesellschaft für linguistische Datenverarbeitung 2001, S. 93-103. [http://www.uni-giessen.de/fb09/ascl/gldv2001/]

Rehm, Georg: E-Mail-ähnliche Textstrukturen in studentischen Homepages. Unveröffentlichtes Manuskript, Ausarbeitung eines Vortrags auf dem Germanistentag 2001, 30.09.-03.10.2001, Erlangen 2002a.

Rehm, Georg: Towards Automatic Web Genre Identification – A Corpus-Based Approach in the Domain of Academia by Example of the Academic's Personal Homepage. In: Proceedings of the 35th Hawaii International Conference on System Sciences (HICSS-35), Big Island, Hawaii 2002b.

Rehm, Georg: Schriftliche Mündlichkeit in der Sprache des World Wide Web. In: *Ziegler, Arne und Christa Dürscheid* (Hg.): Kommunikationsform E-Mail. Tübingen: Stauffenburg 2003, S. 263-308.

Roussinov, Dmitri, Kevin Crowston, Mike Nilan, Barbara Kwasnik, Jin Cai und Xiaoyong Liu: Genre Based Navigation on the Web. In: Proceedings of the 34th Hawaii International Conference on Systems Sciences (HICSS-34), IEEE 2001.

Runkehl, Jens, Peter Schlobinski und Torsten Siever: Sprache und Kommunikation im Internet – Überblick und Analysen. Opladen, Wiesbaden: Westdeutscher Verlag 1998.

Shepherd, Michael und Carolyn Watters: The Functionality Attribute of Cybergenres. In: Proceedings of the 32nd Hawaii International Conference on Systems Sciences (HICSS-32), IEEE 1999.

Stamatatos, Efstathios, Nikos Fakotakis und George Kokkinakis: Automatic Text Categorization in Terms of Genre and Author. In: Computational Linguistics 26 (4), 2001, S. 471-495.

Toms, Elaine G. und D. Grant Campbell: Genre as Interface Metaphor: Exploiting Form and Function in Digital Environments. In: Proceedings of the 32nd Hawaii International Conference on Systems Sciences (HICSS-32), IEEE 1999.