In: *Kommunikationsform E-Mail*, Arne Ziegler, Christa Dürscheid (Hrsg.), Stauffenburg, 2002

Schriftliche Mündlichkeit in der Sprache des World Wide Web

Georg Rehm

Justus-Liebig-Universität Gießen
Angewandte Sprachwissenschaft und Computerlinguistik
Otto-Behaghel-Straße 10 D
35394 Gießen

Telefon: +49 641 99 29052
Fax: +49 641 99 29059

Georg.Rehm@uni-giessen.de
http://www.uni-giessen.de/~g91063/

14. September 2001

Zusammenfassung

Anhand eines großen Korpus von etwa 1,2 Millionen deutschsprachigen HTML-Dokumenten aus der Domäne der akademischen Webserver wird mit Hilfe einer empirischen Studie der Einfluss verschiedener sprachlicher Phänomene, die bzgl. der asynchronen E-Mail-, Newsgruppen-, sowie der synchronen Chat-Kommunikation bereits ausführlich in der Literatur beschrieben wurden, auf Webseiten untersucht. Da sich viele dieser an der konzeptionellen Mündlichkeit orientierenden sprachlichen Phänomene mit computerlinguistischen Methoden erkennen lassen, finden die Analysen vollautomatisch statt und geben Auskunft über Merkmale der konzeptionellen Mündlichkeit im World Wide Web.

Inhaltsverzeichnis

1 Einleitung

Arbeiten zur computervermittelten Kommunikation (Computer-Mediated Communication, CMC) beziehen sich im deutschsprachigen Raum beinahe ausschließlich auf diejenigen Kommunikationsdienste des Internet, in denen – synchron oder asynchron – zwei oder mehr Personen direkt miteinander kommunizieren: Mittlerweile wurden zahlreiche Beiträge vorgelegt, die sich diesbezüglich mit der elektronischen Post (E-Mail), den Newsgruppen des Usenet, dem Internet Relay Chat (IRC) und Gesprächsrunden im World Wide Web (Web-Chats) beschäftigen.

Angesichts der Größe des World Wide Web (Berners-Lee et al., 1992) von weltweit mehr als einer Milliarde Dokumenten, die von über 6,5 Millionen Webservern (`http://www.inktomi.com/webmap/`, Februar 2001) angeboten werden, erscheint es angebracht, dem Schattendasein des World Wide Web in der deutschsprachigen CMC-Forschung entgegen zu wirken. Basierend auf den zahlreichen sprachlichen Phänomenen, die bislang in der Literatur zur computervermittelten Kommunikation berichtet wurden, versuche ich, mit empirischen Methoden, die auf einem Korpus von mehr als einer Million deutschsprachigen HTML-Dokumenten beruhen, eine intuitiv erscheinende Hypothese zu überprüfen.[1]

Viele Internet-Benutzer erstellen persönliche Homepages, auf denen sie sich selbst, ihren Beruf oder Studiengang, private Interessen, Hobbies und viele weitere Informationen einer weltweiten Öffentlichkeit präsentieren. Der Schritt, eine solche Selbstdarstellung im World Wide Web zu publizieren, erfolgt sicherlich erst, nachdem man den ältesten und grundlegendsten Kommunikationsdienst des Internet, E-Mail – und darüber hinaus evtl. Newsgruppen oder das IRC – und deren medienspezifische Besonderheiten kennengelernt hat; diese (para)sprachlichen Spezifika markieren häufig eine konzeptionelle Mündlichkeit (Koch und Oesterreicher, 1994) und sind bereits unter den verschiedensten Gesichtspunkten analysiert worden. Aus den genannten Prämissen lässt sich die folgende Arbeitshypothese formulieren:

> Noch nicht mit einer persönlichen Homepage im World Wide Web vertretene Internet-Benutzer, die durch (grapho)stilistische Elemente wie beispielsweise Smileys, Slangausdrücke sowie isoliert vorkommende, markierte Verbstämme und die hierdurch erzeugte, konzeptionell oftmals primär mündliche Schriftlichkeit der verschiedenen Kommunikationsdienste des Internet geradezu „überflutet“ wurden, fügen derartige – für sie häufig neue und daher evtl. faszinierende – sprachliche Elemente auch in ihre persönlichen Homepages ein, um das Beherrschen des Meta-Mediums Internet, eine Kenntnis der sprachlichen Konventionen und Codes und somit eine Zugehörigkeit zur Gruppe der Netizens zu demonstrieren.

Eine derartige Hypothese lässt sich nur mit empirischen Methoden verifizieren. Das zu diesem Zweck herangezogene Korpus deutschsprachiger HTML-Dokumente umfasst mehr als eine Million Dokumente, die von den Webservern 25 verschiedener deutscher Universitäten stammen.

Abschnitt 2 geht auf die bisherigen, für die vorliegende Studie relevanten Arbeiten zur computervermittelten Kommunikation ein, wobei eine Sammlung derjenigen sprachlichen Merkmale und Phänomene aufgestellt wird, die der konzeptionellen Mündlichkeit (Koch und Oesterreicher, 1994) nahestehen. Das Vorkommen eben dieser Merkmale im World Wide Web – speziell in persönlichen Homepages von Studentinnen und Studenten – wird in dem vorliegenden Beitrag untersucht. Abschnitt 3 beschreibt das eingesetzte Korpus von HTML-Dokumenten hinsichtlich Umfang (3.1), enthaltenen E-Mail Dateien (3.2) und Zugriff der Daten (3.3). Abschnitt 4 stellt den Hauptteil dieses Beitrags dar, die Analyse der Stichproben bezüglich verschiedener sprachlicher Merkmale für konzeptionelle Mündlichkeit.

[1] Der Autor bedankt sich bei Maja Bärenfänger, Petra S. Bayerl, Henning Lobin und Katherine J. Williams (Gießen) sowie Alexander Krumeich (Hamburg) für wertvolle Anregungen.

2 Konzeptionelle Mündlichkeit in computervermittelter Kommunikation (CMC)

Koch und Oesterreicher (1994) differenzieren zwischen medialer und konzeptioneller Mündlichkeit bzw. Schriftlichkeit. Der Aspekt des Mediums bezieht sich hierbei lediglich auf die Realisierung einer sprachlichen Äußerung, die entweder mündlich (d. h. phonisch, z. B. ein Gespräch oder Vortrag) oder schriftlich (d. h. graphisch, z. B. ein Brief, Kochrezept oder Memo) stattfinden kann. Der Aspekt der Konzeption hingegen ist nicht dichotomisch, sondern wird durch ein Kontinuum repräsentiert, dessen Pole einerseits „konzeptionelle Mündlichkeit", andererseits „konzeptionelle Schriftlichkeit" darstellen. Die Konzeption einer sprachlichen Handlung bezieht sich auf „den Duktus, die Modalität der Äußerung" (Storrer, 2000a, S. 153), so wird beispielsweise ein wissenschaftlicher Vortrag zwar medial mündlich realisiert, er steht jedoch der konzeptionellen Schriftlichkeit sehr nahe. Ein Gespräch unter Freunden wird ebenfalls medial mündlich realisiert, dieses steht jedoch der konzeptionellen Mündlichkeit näher. Koch und Oesterreicher (1994) zeigen, dass sich die beiden Endpunkte dieses Kontinuums durch Parameter wie raum-zeitliche Nähe (tendiert zur Mündlichkeit) bzw. Distanz (tendiert zur Schriftlichkeit) beschreiben lassen, wobei skalare Merkmale wie etwa Emotionalität, Vertrautheit der Kommunikationspartner, Situations- und Handlungseinbindung, Spontaneität, Dialog/Monolog und Themenfixierung Einfluss darauf haben, wo eine konkrete sprachliche Äußerung auf diesem Kontinuum anzusiedeln ist. Die meisten Arbeiten, auf die im Folgenden eingegangen wird, diskutieren die per E-Mail, Usenet oder IRC (speziell hierzu Beißwenger, 2001) stattfindende Internet-Kommunikation anhand der medialen bzw. konzeptionellen Mündlichkeit und Schriftlichkeit.

Einer der ersten im deutschsprachigen Raum verfassten Beiträge zur computervermittelten Kommunikation, Lenke und Schmitz (1995), untersucht die Dienste E-Mail, Usenet und insbesondere den Internet Relay Chat (IRC): Neben einer allgemein informellen, „kollegiale[n] und zwanglosen Form" (S. 118) der E-Mail Kommunikation werden Smileys und verbale Beschreibungen realer und fiktiver Handlungen als Mittel eingesetzt, um das Fehlen non-verbaler Signale zu kompensieren.[2] Lenke und Schmitz gehen vornehmlich auf die vielfältigen sprachlichen Phänomene des IRC ein, die u. a. das Turn-Taking, intertextuelle und intermediale Rollenspiele sowie ritualisierte Begrüßungen und Verabschiedungen betreffen.

Feldweg et al. (1995) fokussieren auf den Sprachgebrauch in deutschsprachigen Newsgruppen. Anhand von Frequenzanalysen des mehr als 430 000 Beiträge umfassenden 1993'er Jahrgangs fast aller deutschen Newsgruppen, die mit einem Korpus der *Frankfurter Rundschau* kontrastiert werden, weisen sie hochfrequente Wörter nach (wie z. B. „ich", „man", „du", „mal", „einfach", „ziemlich", „irgendwie" etc.), die den deutlichen Einfluss gesprochener Sprache auf dieses Kommunikationsmedium kennzeichnen.

Günther und Wyss (1996) untersuchen ein Korpus von in der Schweiz produzierten E-Mails und finden hierbei verschiedene sprachliche Phänomene, die sie als „Elemente der Mündlichkeit" bezeichnen, ohne jedoch dabei Bezug auf Koch und Oesterreicher zu nehmen. Die Autorinnen berichten Regionalismen und dialektale Ausdrücke, produktionsbedingte Normabweichungen, Dialogizität, insgesamt sehr kurze Texte und eine „Bildlichkeit" (S. 75), die sich durch den Einsatz von Smileys (Emoticons), individuellen Formatierungen, ASCII-Art und durch die aus dem Internet- und UNIX-Jargon bekannten, oft humorvollen Abkürzungen (ROTFL, IMHO etc.) konstituiert.

Quasthoff (1997) untersucht die Mailing-Listen LINGUIST-List und ETHNO-List (siehe auch Gruber, 1997) und entdeckt hierbei u. a. viele orthographische Fehler, durchgängige Kleinschrei-

[2] In einem Ratgeber zum effektiven Einsatz des Mediums E-Mail wurden Smileys bereits 1994 erläutert als: „Too often the lack of inflection or facial expression can cause a typed phrase in an e-mail message to be interpreted incorrectly. A visual shorthand using *smileys* or *emoticons* has emerged to help the reader decipher the writer's original intent. Smileys are the equivalent of e-mail slang and should not be used in formal business e-mail messages. Also keep in mind that overuse of smileys marks you as a beginner." (Angell und Heslop, 1994, S. 111)

bung, Abkürzungen und „tageszeitorientierte Grußformel[n]“ (S. 42). Durch die hierdurch hervorgerufene „Flüchtigkeit der Botschaft“ wird „eher der Rahmen einer schnell hingeworfenen Notiz als eines Briefes erzeugt.“ (S. 42).

Haase et al. (1997) verknüpfen die Untersuchung von E-Mail, Usenet und IRC mit Koch und Oesterreicher (1994) und kommen zu dem Schluss, dass in allen, inhärent medial schriftlichen Internet-Kommunikationsformen verschiedenste Merkmale konzeptioneller Mündlichkeit existieren, wobei „mit der sprechsprachliche[n] Konzeption die Sprecher-Hörer-Nähe symbolisch erhöht werden soll“ (Haase et al., 1997, S. 81). Diskutiert werden u. a.: Ideogramme (Smileys), Zustands- und Gefühlsäußerungen mittels prädikativ eingebetteter und evtl. durch Sonderzeichen markierter Verbstämme, Deiktika, emuliertes Flüstern und emulierte Prosodie, die Iterationen von Satzzeichen und einzelnen Buchstaben von Wörtern, Abkürzungen, Akronyme und aus dem Internet- und UNIX-Jargon stammenden Slangausdrücke sowie deren Übergeneralisierung.

Pansegrau (1997) beschäftigt sich mit der „Dialogizität und Degrammatikalisierung in E-mails“, wobei u. a. Anredesequenzen und die in der E-Mail-Kommunikation vorhandene erhöhte Toleranz bzgl. Orthographie-, Interpunktions- und Grammatikfehlern untersucht werden. Aufgrund dieser Merkmale sowie der insgesamt feststellbaren sprachlichen Kreativität „wird argumentiert, daß E-mails nicht einen defizitären Stil, sondern eine zweckmäßige und kreative Anpassung an veränderte Kommunikationskanäle repräsentieren“ (S. 95).

Sehr umfangreiche Untersuchungen bzgl. der E-Mail-, Usenet- und IRC-Kommunikation befinden sich in Runkehl et al. (1998). Die Autoren untersuchen anhand verschiedener Korpora Merkmale wie durchgängige Kleinschreibung, Bigraphen, unterschiedliche Typen von Fehlern, Akronyme, Smileys, Assimilationen, Reduktionen, Iterationen, Anreden und Verabschiedungen, Signaturen, Dialektismen, Diskurspartikeln und Interjektionen: „Je stärker die Kommunikation dialogischer und synchroner erfolgt [E-Mail → Usenet → Chat, G. R.], desto häufiger lassen sich mündliche Aspekte des Sprachgebrauchs in der Internet-Kommunikation feststellen.“ (S. 116).

Grzega (1999) untersucht empirisch auf der Grundlage traditioneller und elektronischer Briefe sowie von Fragebögen den Aspekt der Differenz zwischen herkömmlicher und digitaler Post sowie deren gegenseitige Beeinflussung. Er betont, dass *kein* genereller „e-style“ (S. 15) existiert: „The boundaries between formality and informality (private letter vs. business letter) appear much more fuzzy, but it seems entirely unjustified to speak of overall present features. [... M]ore empirical studies are needed in the research of e-style.“ (S. 16).

Dürscheid (1999) untersucht, „welche sprachlichen Merkmale kennzeichnend für die Internetkommunikation sind“, wobei Chat, E-Mail und Usenet betrachtet werden. Neben der Zuordnung von Smileys, unflektierten Verbformen (oftmals Verbletztkonstruktionen), orthographischen Fehlern und Interjektionen stellt sie fest, dass auch wiederholte Ausrufe- und Fragezeichen sowie iterierte Buchstaben und konstante Großschreibung Merkmale für konzeptionelle Mündlichkeit sind, wobei die Authentizität der sprachlichen Handlung oft im Vordergrund steht: „Das Motto [im Chat, G. R.] scheint zu sein: ‘Schreib, wie Du sprichst’ und ‘Schreib so schnell, wie Du kannst’.“ (S. 21).

Storrer (2000a, S. 153 f.) fasst die für die Internet-Kommunikation wesentlichen Charakteristika konzeptioneller Mündlichkeit zusammen: Bezüglich der Lexik herrscht offenbar eine Präferenz für einfache und kurze Wörter, sowie umgangssprachlich markierte und dialektale Ausdrücke vor. Auf der syntaktischen Ebene findet sich häufig ein parataktischer, tw. fehlerhafter Satzbau, der oftmals typisch sprechsprachliche Konstruktionen aufweist. Weiterhin ist für Storrer eine „freie, assoziative, dialogisch gesteuerte Themenentwicklung“ (S. 154) charakteristisch:

> Die kommunikative Grundhaltung der Mündlichkeit orientiert sich an dem Setting des alltäglichen Gesprächs von Angesicht zu Angesicht zwischen miteinander vertrauten Gesprächspartnern, die sich in der Sprecher- und Hörerrolle abwechseln. Typisch für dieses Setting sind kurze Planungszeiten bei der Produktion und kurze Verarbeitungszeiten für die Rezeption. Die Äußerungen werden meist spontan gebildet; die Themenentwick-

lung ist offen, wobei für die Teilnehmer in der Hörerrolle stets die Option der Rückfrage oder des Einspruchs besteht und die Teilnehmer in der Sprecherrolle mit sprachlichen und mimisch-gestischen Mitteln Feedback erhalten.

Betrachten wir nun das World Wide Web als potentielles Untersuchungsobjekt der CMC-Forschung, so liegen nur wenige Arbeiten vor, die sich mit diesem Thema auseinander setzen.[3] Dürscheid (2000) fokussiert auf dem Hypertext-Aspekt und geht auf Unterschiede zwischen Webseiten und gedruckten Texten ein. Zentral ist hierbei die These, dass Webseiten – Dürscheid meint hier explizit nur diejenigen Webseiten, die Teil eines Hypertextes sind, der mehrere Knoten umfasst – ausschließlich durch „die Rekurrenz einzelner Textteile" sprachlich miteinander verbunden werden können: „Andere anaphorische Mittel, insbesondere solche, die der Textverdichtung dienen, wie z. B. die Verwendung von Pronomina [...], kann es nicht geben. Eben weil Webseiten von verschiedenen Richtungen aus anwählbar sind, wären solche Ausdrucksformen nicht interpretierbar" (S. 66). Bezüglich der strikten medialen Trennung – *entweder* schriftlich *oder* mündlich – innerhalb des von Koch und Oesterreicher (1994) vorgeschlagenen Modells meint Dürscheid: „Eine Webseite, die aus multimedialen Elementen besteht, ist nicht mehr kanonisch einer der beiden Repräsentationsformen, mündlich oder schriftlich, zuzuordnen, denn sie macht Gebrauch von mehreren Zeichenträgern zur selben Zeit."

Diekmannshenke (2000) untersucht im World Wide Web verfügbare Gästebücher, vergleicht diese mit der papierenen Variante, diskutiert die neuen Ausprägungen der Konzepte „Gast" und „Gastgeber" und stellt fest: „Sprachliche Kenntnisse sind für diese Art der Kommunikation nur in geringem Maße erforderlich, handelt es sich doch vielmehr nur um das Hinterlassen einer schriftlichen 'Besuchsspur'." (S. 138). Diekmannshenke zeigt anhand vieler Belege, dass in Gästebüchern auch diejenigen der konzeptionellen Mündlichkeit nahestehenden sprachlichen Merkmale eingesetzt werden, die bislang für die E-Mail-, Chat- und Newsgruppen-Kommunikation berichtet wurden, z. B. Verbstämme, Jargonausdrücke und Abkürzungen, jedoch „werden diese Mittel bei Gästebucheinträgen aber nur in geringem Umfang verwendet" (S. 141).

Storrer geht in verschiedenen Artikeln (1999b, 2000b) speziell auf das World Wide Web ein: Storrer (1999b) untersucht Homepages mit Hilfe der Aspekte „Hypertext", „Multimedia", „Interaktivität" und „computervermittelte Kommunikation". Bezüglich des letzten Punktes sieht Storrer den großen Vorteil des Webs in der Verbindung von Information und Kommunikation, da Benutzer nicht nur Informationen abrufen können, sondern mit Hilfe der in einen Browser integrierten Funktionalität auch die Kommunikationsdienste des Internet in Anspruch nehmen können. Storrer (2000b) geht vor allem auf theoretische Aspekte von Hypertext ein und beantwortet die Frage, ob für Hypertext ein neuer Textbegriff notwendig sei, mit einem entschiedenen Nein. Für den vorliegenden Beitrag ist ihre an Koch und Oesterreicher (1994) angelehnte Unterscheidung zwischen medialer Linearität bzw. Nicht-Linearität und konzeptioneller Linearität bzw. Nicht-Linearität wichtig. Bezüglich des Mediums herrscht Linearität vor, wenn die Rezeption der Daten fest vorgegeben ist (z. B. Tonband, Videokassette); Nicht-Linearität herrscht in Bezug auf das Medium, wenn Daten in unterschiedlicher Abfolge rezipiert werden können. Bezüglich der Konzeption meint Storrer: „Konzeptionelle Linearität bzw. konzeptionelle Nicht-Linearität sind Eigenschaften, die sich auf die vom Textproduzenten getroffene Entscheidung für eine der nachfolgend skizzierten Strukturierungsformen beziehen [monosequenzierte Texte, mehrfachsequenzierte Texte, unsequenzierte Texte, G. R.]." Gerade die beiden letzten Ausprägungen entsprechen hierbei bewusst vom Autor aufgebrochenen Texten, die nicht mehr in fest vorgegebener Reihenfolge rezipiert werden müssen. Stattdessen können Leser denjenigen Leseweg auswählen, der ihnen am geeignetsten erscheint. Unsequenzierte Texte

[3] Sprachwissenschaftliche Arbeiten beschäftigen sich diesbezüglich fast ausschließlich mit sehr abstrakten, oftmals nicht empirisch untersuchten Aspekten der Textkohärenz in Hypertexten, vgl. beispielsweise van Berkel und de Jong (1999), Storrer (1999a), Fritz (1999) oder etwa digitalen Ausgaben von Zeitungen (Bucher, 1996).

(etwa Lexika oder Wörterbücher) schließlich geben nicht einmal einen Leseweg vor, sie stellen extensiv vernetzte Sammlungen einzelner, in sich kohärenter Informationsknoten dar.

3 Hypnotic – Ein Korpus von HTML-Dokumenten

Im Rahmen des Projekts *Hypnotic*[4] entsteht ein Korpus von HTML-Dokumenten. Das Projekt beschäftigt sich mit der Untersuchung konventionalisierter Textstrukturen – *Hypertextsorten* – und entwickelt Methoden zur maschinellen Klassifikation von HTML-Dokumenten in ihre jeweiligen Hypertextsorten, um auf diese Weise Suchmaschinen mit erweiterter Funktionalität auszustatten, so dass deren Benutzer effizienter und gezielter Dokumente auffinden können (Rehm, 2002).

Da eine vollständige empirische Untersuchung aller im WWW existenten Hypertextsorten den Rahmen des Projektes sprengte, wurden zwei wesentliche Einschränkungen bei der Sammlung der im Korpus enthaltenen Dokumente vorgenommen: Zum einen beschränke ich mich auf die *Webserver deutscher Universitäten*, zum anderen sind lediglich *deutschsprachige Dokumente* von Interesse.

Die wichtigsten Komponenten, die die automatische Datensammlung vornehmen, sind ein umfangreich konfigurierbarer Web-Roboter sowie ein Werkzeug zur Sprachidentifizierung. Dieses überprüft mit Hilfe verschiedener statistischer Metriken für jedes heruntergeladene Dokument, ob es in Deutsch geschrieben wurde. Ist dies der Fall, wird das Dokument in die Korpusdatenbank aufgenommen. Technische Details zur Datensammlung, zur Implementierung der angesprochenen Beschränkungen sowie zum Datenbankzugriff befinden sich in Rehm (2001).

3.1 Umfang des Korpus

Derzeit (Ende August 2001) befinden sich 1 137 071 HTML-Dokumente, die auf 3 632 verschiedenen Webservern von 25 verschiedenen Universitäten[5] angeboten werden, in der Hypnotic-Korpusdatenbank. Neben diesen etwa 12 Gigabyte Daten sind weitere Dateitypen in der Datenbank enthalten: 7 258 Cascading Style Sheets (CSS-Dateien), 16 510 XML-Dateien, 663 SGML-Dateien, 79 464 ASCII-Dateien, 254 News-Artikel und 125 E-Mails. Benutzt man diese Daten als Grundlage, ist abschätzbar, dass die endgültige Version des Korpus etwa 3 000 000 deutschsprachige Webseiten von etwa 60 Hochschulen beinhalten wird.

3.2 E-Mail Dateien

Die Tatsache, dass E-Mails mit einem eigenständigen Dateityp im World Wide Web angeboten werden, erscheint auf den ersten Blick sehr zu verwundern, hängt jedoch mit dem internen Verfahren des Hypertext Transfer Protocols (HTTP, siehe Fielding et al., 1999) zusammen, das den Transfer von Dateien zwischen Webserver und Browser spezifiziert: Ein HTML-Dokument muss vom Browser als solches identifiziert werden, damit es mitsamt der enthaltenen Grafiken und Photos entsprechend dargestellt werden kann. Dies geschieht nicht über den Dateisuffix (üblicherweise `.html` oder `.htm`), sondern über den HTTP-Header, der – für den Benutzer unsichtbar – vom Server zum Browser geschickt wird. Hierin wird angegeben, ob es sich bei einer Datei um HTML-Text (`text/html`), eine Grafik (z. B. `image/gif` oder `image/jpeg`) oder eine PDF-Datei (`application/pdf`) handelt. Je nach Dokumenttyp kann der Browser nun das Dokument entweder selbst darstellen oder, wie etwa im Falle eines PDF-Dokuments, ein externes Programm starten.

[4] „Hypertexts and their Organisation into a Taxonomy by Means of Intelligent Classification“, siehe `http://hypnotic.germanistik.uni-giessen.de`

[5] Diese sind im Einzelnen die Universitäten Augsburg, Düsseldorf, Duisburg, Erlangen-Nürnberg, Essen, Gießen, Heidelberg, Koblenz, Leipzig, Paderborn, Passau, Trier, Ulm und Witten-Herdecke, die Technischen Universitäten Chemnitz, Cottbus, Dresden, Freiberg und Hamburg-Harburg, die RWTH Aachen, Fernuniversität Hagen, FU Berlin, HDK Berlin, Katholische Universität Eichstätt sowie die European Business School in Oestrich-Winkel.

Ein bestimmter derartiger Dokumenttyp lautet `message/rfc822` und wird zur Identifizierung von E-Mails eingesetzt.[6] Da sich der vorliegende Band mit der „Kommunikationsform E-Mail“ beschäftigt, erscheint eine nähere Untersuchung dieser Dateien angebracht: Fast alle der 125 Dateien vom Typ `message/rfc822` sind Teil online zugänglicher Archive oder Digests (monatliche Aufstellungen) öffentlicher oder interner Mailing-Listen[7] und liegen im `.mbox`-Format vor, d. h. in der Datei befindet sich erst der Header einer Mail, daraufhin deren Rumpf; getrennt durch eine Leerzeile folgt optional eine weitere Mail, wieder getrennt in Header und Rumpf usw., so dass die 125 Dateien insgesamt mehr als 1 000 E-Mails enthalten. Andere E-Mail Dateien stammen von persönlichen Homepages und enthalten Informationen, die beispielsweise von rein privatem Interesse sind (in einem Verzeichnis namens `todo`), Software-Ankündigungen oder Digests einer Oscar Wilde Mailing-Liste (siehe die Verweise auf `http://www.math.fu-berlin.de/~guckes/wilde/`).

3.3 Datenzugriff

Der Zugriff auf die im Korpus enthaltenen Dokumente erfolgt mit Hilfe einer eigens in der Skriptsprache PHP implementierten WWW-Schnittstelle, die auf die zugrundeliegende SQL-Datenbank zugreift, die Metadaten von allen im Korpus verfügbaren Dateien enthält. Mit Hilfe dieser Schnittstelle kann sich ein Benutzer die Server einzelner Universitäten anzeigen lassen und daraufhin einen Server auswählen, um ein Dokument genauer zu untersuchen (Abbildung 1, S. 26, zeigt den Modus der Dokumentansicht der Oberfläche). Alternativ kann auch korpusweit nach Webservern gesucht werden, deren Namen einem bestimmtem Muster (z. B. `informatik` oder `linguistik`) entsprechen. Ähnliche Suchmöglichkeiten existieren für Dokumente.

Neben diesen direkten Zugriffsmethoden verfügt das System über eine komfortable Maske zur zufallsbasierten Generierung von Stichproben beliebiger Größe. So kann man beispielsweise eine Stichprobe generieren lassen, die 100 Dokumente enthält, die eine Länge von mindestens 5 000 Bytes und, bezogen auf ihre jeweilige Adresse, eine „Tiefe“ von mindestens sechs Verzeichnisebenen besitzen, aus den Universitäten Leipzig, Dresden, Gießen und Heidelberg stammen, und einem bestimmten Suchmuster (beispielsweise `kontakt`, `bericht` oder `telefon`) entsprechen. Die auf diese Weise zufällig aus der Korpusdatenbank extrahierte Liste von Dokumenten lässt sich daraufhin abspeichern, damit sie für einen späteren Zugriff erneut verfügbar ist. Dieser Zugriff kann bei Bedarf auf bestimmte Benutzer der WWW-Schnittstelle des Hypnotic-Systems eingeschränkt werden.

3.4 Drei unterschiedliche Stichproben

Für diesen Beitrag wurden drei Stichproben[8] gezogen, die zu unterschiedlichen thematischen Bereichen gehören:

S1 *Persönliche Homepages von Studierenden*: 25 481 HTML-Dokumente, die von manuell ausgewählten Webservern stammen und Adressen besitzen, die ausschließlich auf studentische Homepages zutreffen. Hierzu gehören etwa verschiedene Webserver von Studentenwohnheimen[9] oder die an der Universität Gießen eindeutig identifizierten studentischen Homepages. Diese

[6] Die Abkürzung RFC822 steht für das *Request for Comments* Dokument (RFC) mit der Nummer 822. RFCs definieren diejenigen Protokolle, auf denen die gesamte technische Realisierung des Internet, wie beispielsweise auch das o. a. HTTP-Protokoll (RFC 2616), basiert. RFC 822 (Crocker, 1982) definiert das Format von E-Mails und wurde im April 2001 durch eine überarbeitete Spezifikation, RFC 2822, abgelöst.

[7] Die im Korpus enthaltenen Dateien des Webservers der Professur für Betriebssysteme (TU Dresden) stellen beispielsweise WWW-Archive der Mailing-Listen einzelner Lehrveranstaltungen dar, vgl. `http://os.inf.tu-dresden.de/mailman/listinfo/`.

[8] Vollständige Listen der Adressen der in den drei Stichproben enthaltenen Dokumente sowie das jeweilige Datum des Downloads befinden sich unter `http://www.uni-giessen.de/~g91063/html/homepage-analysis/`.

[9] Beispielsweise `www.wohnheim.uni-ulm.de` oder `www.kawo1.rwth-aachen.de`

befinden sich auf dem Server `wwwstud.uni-giessen.de`, und ihre Pfadkomponente beginnt jeweils mit `/~s`.

S2 *Persönliche Homepages von Mitarbeitern*: 14 247 HTML-Dokumente, die ebenfalls von manuell ausgewählten Webservern stammen. So bieten z. B. die Server `www.physik.uni-ausgburg.de`, `www.linguistik.uni-erlangen.de` oder `www.imise.uni-leipzig.de` auf dem Adressmuster, das der üblichen Konvention persönlicher Dokumente entspricht (`http://www.../~...`), mit nur sehr wenigen Ausnahmen die persönlichen Homepages von Mitarbeitern der jeweiligen Universitäten bzw. Forschungseinrichtungen an.

S3 *Tief eingebettete Dokumente*: 10 000 mit Hilfe der in Abschnitt 3.3 angesprochenen Maske von der Korpusdatenbank zufällig ausgewählte HTML-Dokumente, die eine Tiefe von mindestens drei und maximal zehn Verzeichnissen[10] besitzen und deren jeweilige Pfadkomponente nicht mit `/~` beginnt. Durch diese Einschränkungen sollte verhindert werden, dass sich persönliche Dokumente in diesem dritten Sample befinden. Stattdessen enthält diese Stichprobe größtenteils HTML-Versionen von technischen Berichten oder Dokumente administrativer Natur.

4 Analyse der Stichproben

Die im Folgenden dargestellte maschinelle Untersuchung der drei im vorherigen Abschnitt beschriebenen Stichproben erfolgte mit Hilfe eines in der Programmiersprache Perl (Wall et al., 2000) implementierten Skripts, das das Modul `HTML::Parser` einsetzt, um einen effizienten Zugriff auf HTML-Elemente (Hypertext Markup Language, Raggett et al., 1999), deren Attribute und den eigentlichen Inhalt von Elementen – den auf Webseiten enthaltenen Text – zu ermöglichen.

4.1 Umfang der Stichproben und HTML-Merkmale

Zunächst werden der Umfang der drei Stichproben sowie grundlegende Merkmale des jeweiligen HTML-Codes der enthaltenen Dokumente betrachtet (vgl. Tabelle 1). Sehr deutlich fällt auf, dass bzgl. der Anzahl Token[11] pro Dokument nur sehr geringe Schwankungen zwischen den Stichproben existieren, wobei die Homepages der Mitarbeiter – sowohl durchschnittlich als auch im Median – umfangreicher sind als die Dokumente aus S3.

Bezüglich des Einsatzes von Framesets zur visuellen Strukturierung von Dokumenten bildet S1 diejenige Stichprobe mit den prozentual häufigsten Vorkommen. Viele Bücher, die sich mit Web Design beschäftigen, raten aus den verschiedensten Gründen vom Gebrauch von Framesets ab. Diese einfach zu realisierende Strukturierungshilfe wird im eher professionellen Bereich (S2, S3) nicht so häufig eingesetzt. Ein Dokument, das ein `<frameset>` Element enthält, referenziert verschiedene Dateien, die die tatsächlichen Inhalte einer Webseite – den Inhalt der einzelnen Frames – enthalten. Da Dateien, die über ein `<frameset>` Element verfügen, fast ausnahmslos keinen sprachlichen Inhalt besitzen, wurden diese bei der Auswertung der Daten nicht berücksichtigt.

Betrachtet man die jeweiligen Daten bzgl. des Einsatzes von eingebetten Bildern (diese umfassen alle Vorkommen des HTML-Elements `<img>`), fallen nur wenige Unterschiede zwischen den drei Stichproben auf. Die Vermutung, dass bei der visuellen Gestaltung studentischer Homepages sehr extensiv mit dem Einsatz von Graphiken und Photos umgegangen wird, kann hier nicht bestätigt

[10] Beispielsweise `http://www.tu-harburg.de/v/studinf/mb/hs_eng/bruchme.htm`, der Kommentar der Lehrveranstaltung „Bruchmechanik und Schwingfestigkeit“ des Studiengangs Maschinenbau an der TU Hamburg-Harburg.

[11] Ein Token ist in dieser Untersuchung definiert als eine aus beliebigen Zeichen (außer Zwischenraum, d. h. Leerzeichen, Tabulatorzeichen oder Zeilenwechsel) bestehende Zeichenkette (üblicherweise ein Wort, jedoch subsumiert dieser Begriff auch Abkürzungen, Zahlen, Ketten von Sonderzeichen etc.), die eine Länge von einem Zeichen oder mehr umfasst. Als regulärer Ausdruck (vgl. etwa Friedl, 1997) in Perl: `[^\s]{1,}`.

werden; ganz im Gegenteil, der arithmetische Durchschnitt und der Mittelwert der eingebetteten Bilder pro Dokument ist in S3 sogar höher als in S1.

Auch bzgl. der Verknüpfung ähneln sich die drei Stichproben sehr: Zwischen 81,6% (S1) und 89% (S3) der Dokumente besitzen mindestens einen Hyperlink.[12] In allen drei Samples sind durchschnittlich etwa 12 Hyperlinks pro Dokument enthalten (Median jeweils 6).[13] Die Inhalte der jeweiligen Hyperlink-Anker – die typographisch hervorgehobenen Wörter, deren Aktivierung den Benutzer zu einem neuen Dokument führt – werden in Abschnitt 4.3 genauer untersucht.

Merkmal	Studentische Homepages (S1)	Mitarbeiter-Homepages (S2)	Tiefe Dokumente (S3)
Anzahl Dokumente	25 481	14 247	10 000
Anzahl Token	8 531 088	4 968 339	3 125 195
Token pro Dokument	ø367 (Med.: 94)	ø373 (Med.: 114)	ø321 (Med.: 91)
Min./Max. Anzahl Token	1 / 53 203	1 / 28 280	1 / 32 341
HTML-Deklaration	8 958 (35,1%)	5 980 (42%)	3 240 (32,4%)
Enthält Frameset	1 075 (4,2%)	538 (3,8%)	293 (2,93%)
Gesamtanzahl Bilder	116 617	49 969	53 948
Mind. ein Bild	15 369 (63%)	8 967 (65,4%)	6 477 (66,7%)
Min./Max. Anzahl Bilder	0 / 617	0 / 199	0 / 219
Bilder pro Dokument	ø7,6 (Med.: 3)	ø5,6 (Med.: 3)	ø8,3 (Med.: 5)
Bilder als Hyperlinks	40 251	22 199	26 058
Gesamtanzahl Hyperlinks	257 047	136 074	96 777
Mind. ein Hyperlink	19 906 (81,6%)	11 634 (84,9%)	8 641 (89%)
Min./Max. Anzahl Links	0 / 951	0 / 1 544	0 / 473
Hyperlinks pro Dokument	ø12,9 (Med.: 6)	ø11,7 (Med.: 6)	ø11,2 (Med.: 6)
Wörter pro Hyperlink	ø2 (Med.: 1)	ø2,3 (Med.: 2)	ø2,1 (Med.: 1)

Tabelle 1: Umfang und HTML-Merkmale der drei Stichproben

4.2 Wortfrequenzen

Tabelle 2 zeigt die je 50 häufigsten deutschsprachigen Token der drei Stichproben. Alle englischsprachigen Begriffe, nur einen Buchstaben umfassende Zeichenketten (z. B. „a“) sowie nicht druckbare

[12] Dass zwischen 10% und 20% der Dokumente keinen einzigen Hyperlink enthalten, ist m. E. auf zwei Umstände zurückzuführen: Einerseits stellen viele Dokumente Blätter (im graphentheoretischen Hypertext-Sinn) dar, d. h. die in ihnen enthaltene Information ist beispielsweise ein Gedicht, eine Geschichte, eine Vorlesungsankündigung oder dergleichen, die auch ohne Hyperlink les- und benutzbar ist, wenngleich die Navigationsmöglichkeiten hierunter merklich leiden. Andererseits stellen, gerade in S1 und S2, viele Einstiegsseiten persönlicher Homepages nur „Dummy“-Dokumente ohne jeglichen Hyperlink dar, die von Mitarbeitern des jeweiligen Rechenzentrums angelegt wurden, damit der Webserver beim Aufruf der Adresse keine Fehlermeldung liefert.

[13] Amitay (2000) untersucht etwa 1 000 HTML-Dokumente, vornehmlich persönliche Homepages, und kommt zu vergleichbaren Resultaten: Durchschnittlich enthalten die Dokumente ihrer beiden Korpora 35,9 (Median: 14–15) bzw. 17,7 (Median: 13–14) Hyperlinks: „This consistency is very remarkable. It means that there is a well established convention regarding the length of anchors and their density within a document.“ HTML-Dokumente aus dem englischsprachigen Bereich werden prozentual offenbar mit mehr Hyperlinks versehen, als dies im deutschen Sprachraum üblich ist.

Studentische Homepages (S1)	Mitarbeiter-Homepages (S2)	Tiefe Dokumente (S3)	*tageszeitzung* (1. Halbjahr 1994)
der (191261) und (177014) die (164650) in (100057) von (65855) zu (65397) den (62847) mit (55437) das (51329) für (50370) ist (49443) des (46948) auf (44122) im (43671) nicht (41521) sich (40787) Die (39977) eine (39643) ein (37629) dem (34399) auch (33330) es (30081) ich (28098) an (27399) als (27005) oder (23862) Sie (22278) bei (22140) man (21989) sie (20423) sind (20371) einer (20349) daß (20325) nach (20241) zur (20189) zum (20180) aus (19381) wird (18510) noch (18469) nur (18286) Seite (17764) Der (17513) wie (17323) Das (17183) einen (17075) werden (17024) so (16465) über (16151) aber (15500) um (15347)	der (134341) und (126979) die (91224) in (68281) von (46965) des (40348) für (38137) den (35699) zu (33820) im (31151) mit (30764) das (25568) Die (25254) ist (23619) auf (22889) eine (20591) sich (20193) dem (19609) ein (17062) nicht (16791) an (16601) als (16190) zur (15599) oder (14617) auch (14122) werden (12610) bei (12404) einer (12369) es (11998) zum (11974) durch (11691) wird (11170) aus (10974) sind (10908) Dr. (10092) nach (9824) am (9815) Sie (9780) Der (9607) über (9355) Das (9173) daß (8679) sie (8454) einen (8207) einem (8079) wie (7829) nur (7588) so (7009) noch (6922) um (6921)	der (87638) und (74478) die (62739) in (40783) von (29862) des (25229) den (22923) für (22692) zu (21301) mit (19723) im (18606) Die (16880) das (16526) ist (15738) auf (15088) eine (14393) sich (12452) dem (12270) zur (11818) nicht (11438) ein (10891) als (10354) oder (9883) an (9504) auch (9220) werden (8837) einer (8792) bei (8721) wird (8676) Sie (7327) durch (7232) sind (7103) aus (7056) zum (7056) es (7028) nach (6911) über (6313) Der (6204) Das (5721) einem (5323) daß (5313) einen (5164) wie (5013) nur (4960) sie (4842) Dr. (4741) am (4657) kann (4510) um (4372) eines (4325)	der (251828) die (245264) und (167878) in (130930) den (93742) von (74710) zu (72179) das (67748) mit (62376) sich (61464) nicht (61450) ist (58396) für (54906) auf (54200) des (52838) im (52181) Die (50693) dem (50448) ein (47889) eine (43990) es (38535) als (37569) auch (37093) an (33475) daß (33299) sie (32447) aus (31675) werden (30116) hat (29275) er (27967) nach (26158) noch (25202) Der (24879) einer (24522) wie (23954) wird (23787) sind (23771) um (23398) am (22945) bei (22517) vor (22146) so (21830) nur (21638) Das (21483) über (21391) haben (20880) einem (20080) einen (19642) zum (19378) war (18426)

Tabelle 2: Die je 50 häufigsten deutschsprachigen Token aus S1 – S3 sowie aus dem ersten Halbjahr des Jahrgangs 1994 der *tageszeitung* (enthält 7 654 357 Token). Im Text diskutierte Token wurden speziell markiert.

Sonderzeichen wurden der besseren Lesbarkeit halber aus den Listen entfernt. Zu Vergleichszwecken zeigt die Tabelle auch die 50 häufigsten Token aus dem ersten Halbjahr 1994 der *tageszeitung*.[14]

Vergleicht man die vier Listen, so setzt sich der im vergangenen Abschnitt bereits konstatierte Trend der Ähnlichkeit auf den ersten Blick fort. Etwa die jeweils 20 häufigsten Token sind in allen vier Stichproben praktisch identisch: bestimmte und unbestimmte Artikel, Präpositionen sowie Kon- und Disjunktionen. Bereits an Position 23 findet sich in der Stichprobe der studentischen Homepages das Pronomen „ich“ (S2: 62, S3: 88, taz: 67), das gemeinsam mit dem ebenfalls nur dort vorkommenden „man“[15] (29, S2: 52, S3: 51, taz: 54) einen eher informellen und in der direkten Anrede gehaltenen sprachlichen Stil der persönlichen Homepages der Studierenden andeutet.[16] Dieses Ergebnis bestätigt die Untersuchung von Amitay (2000), die 155 englischsprachige persönliche Homepages analysiert. Dort taucht „I“ bereits an siebter, „you“ an vierzehnter Stelle auf. Zum Vergleich die jeweiligen Positionen von „Du“: S1: 106, S2: 211, S3 und taz: nicht unter den 300 häufigsten Wörtern). Auf studentischen Homepages scheinen die Leser also sehr viel häufiger mit „Du“ angeredet zu werden als auf den Dokumenten der anderen Stichproben.

Neben „ich“ befindet sich ein weiteres Wort, eines der beiden einzigen Nomen in den vorliegenden Listen, in der Liste der 50 häufigsten Token aus S1, „Seite“ (Pos. 41). Dieses wird in den studentischen Homepages synonym für „Webseite“ eingesetzt, wie beispielsweise in `http://www.rzuser.uni-heidelberg.de/~oclanget/linx.htm`: „Linx zu anderen Seiten“. Ausschließlich in S2 (Pos. 35) und S3 (Pos. 46) befindet sich das zweite Nomen, die Abkürzung „Dr.“. Wissenschaftliche Mitarbeiter, Assistenten und Professoren sind beinahe ausschließlich die Autoren der in S2 enthaltenen Dokumente, daher erscheint ein derart häufiges Vorkommen von „Dr.“ dort nicht verwunderlich („Prof.“ befindet sich in S2 an Pos. 53). Die ebenfalls sehr häufigen Vorkommen in S3 sind vermutlich durch Listen der Mitarbeiter von Arbeitsgruppen oder Fachbereichen oder durch Kommentare zu Lehrveranstaltungen zu erklären („Prof.“ in S3: Pos. 73).

In der Liste der 50 häufigsten Wörter der *tageszeitung* sind einige konjugierte Kopulativverben enthalten („hat“, „haben“, „war“), die in S1 – S3 vollständig fehlen. Da Zeitungen über vergangene Ereignisse berichten, verwundert die Präsenz dieser Wörter nicht; die Abwesenheit in den ersten drei Stichproben lässt im Umkehrschluss vermuten, dass Hypertext-Dokumente vornehmlich im Präsens verfasst werden. Amitay (2000) kommt nach einem Vergleich der Frequenzlisten ihres Homepage-Korpus und dem British National Corpus (BNC) zu einem ähnlichen Schluss und vermutet: „This preference is probably due to the fact that hypertext is changeable and that people modify their files whenever the facts change, thus at the time of reading the hypertext document [...] is assumed to be representing a fact.“

4.3 Trigrammfrequenzen in Linkankern

Linkanker sind diejenigen Teile eines HTML-Dokuments, über deren Aktivierung (üblicherweise per Mausklick) der Benutzer zu einem neuen Dokument oder einer anderen Position im aktuellen Dokument gelangt. Eine Untersuchung über die Häufigkeiten der in Linkankern vorkommenden Trigramme – Abfolgen von drei Token – kann evtl. Aufschluss darüber geben, wie Autoren von HTML-Dokumenten den Inhalt des verlinkten Dokuments signalisieren, d. h., welche sprachlichen

[14] Der Autor bedankt sich bei Frank Henrik Müller und Tylman Ule (Tübingen), die freundlicherweise das *taz*-Korpus in einer XML-annotierten Form zur Verfügung gestellt haben.

[15] Feldweg et al. (1995) berichten, dass diese Wörter in einem Korpus von mehr als 430 000 deutschsprachigen Newsartikeln ebenfalls hochfrequent sind.

[16] Hiervon existieren auch extreme Ausprägungen, etwa auf einer studentischen Homepage, die nach der Datensammlung modifiziert wurde, so dass das im Korpus vorliegende Dokument nicht mehr dem Aktuellen entspricht. Hier wird der Leser direkt angesprochen mit „Wilkommen auf meiner Homepage. [...] Wie auch immer, Du hast sie gefunden!“, einige Sätze später werden *alle* potentiellen Leser angesprochen, wie dies auch häufig in Ansagen von Anrufbeantworten zu hören ist: „Auf der linken Seite könnt ihr ein bißchen rumschauen und z.B. ein paar Bildchen ansehen (mit Thumbnails) Natürlich könnt ihr mir auch schreiben!“. Abbildung 1 (Seite 26) zeigt ein weiteres Beispiel dieser gemischten Anrede.

Studentische Homepages (S1)	Mitarbeiter-Homepages (S2)	Tiefe Dokumente (S3)
zurück zur KaWo-Leitseite (758) vorherigen Block laden (754) nächsten Block laden (751) HTML-Dateien selbst erstellen (475) Anfang der Seite (194) Zurück zur Homepage (192) zurück zum Anfang (163) zum Anfang der (150) Traditionelle Chinesische Medizin (136) Zurück zur Startseite (135) für Traditionelle Chinesische (132) Arbeitskreis für Traditionelle (127) zurück zur Übersicht (125) Zurück zur Übersicht (125) der Frankfurter Rundschau (124) Homepage der Frankfurter (122) zur ersten Folie (116) Zurück zur ersten (115) Zurück zu meiner (112) zurück zum Seitenanfang (107) Zurück nach oben (107) zu meiner Homepage (105) Zurück zum Inhaltsverzeichnis (102) AEGEE in Europe (95) AEGEE in Heidelberg (95) Zurück zur Protokolliste (93) Einführung in die (88) Zurück zum Anfang (87) Word-Datei zum Downloaden (82) Die Welt der (76) Welt der Worte (76) Zurück zur Hauptseite (75) Protokoll als Word-Datei (70) eMail an Redaktion (70) als Word-Datei zum (69) Absatztypen und Textgestaltung (64) Big picture index (63) Small picture index (63) der Universität Heidelberg (61) zurück zur Homepage (61)	Zurück zum Anfang (351) zum Anfang des (249) Zurück zur Homepage (246) ein Partizip II (234) Institut für Geographie (224) Mitglied bei page (191) am Institut für (181) Institut für Ernährungswissenschaft (170) Informations- und Dokumentationsstelle (169) der Justus-Liebig-Universität Gießen (166) Dokumentationsstelle am Institut (161) und Dokumentationsstelle am (157) Zurück zur Startseite (156) für Ernährungswissenschaft der (152) Anfang des Dokuments (146) Ernährungswissenschaft der Justus-Liebig-Universität (146) Was ist die (138) ist die GfÖ (135) Mitglieder der GfÖ (135) Zurück zum Inhaltsverzeichnis (134) Zurück nach oben (123) Zurück zur Übersicht (121) in hexadezimaler Angabe (120) Partizip II und (115) Konventionelle syntaktische Analyse (113) der Infinitiv von (109) und der Infinitiv (108) Einführung in die (97) Partizip II von (95) das Partizip II (92) Verlag gesund essen (78) zurück zur Hauptseite (77) Arbeitssicherheit und Umweltschutz (77) Bereich American Football (77) zum Bereich American (76) und ein Partizip (76) Zurück zum Bereich (74) zurück zum Anfang (66) Zurück zur Hauptseite (64) II und der (63)	Zurück zur ersten (662) zur ersten Folie (662) über alle SMT-Bilder (107) Überblick über alle (107) Zurück zur Übersicht (94) German News Team (92) Mail Thread Index (90) Zurück zum Anfang (85) Einführung in die (83) Anfang des Dokuments (70) zum Anfang des (70) Liste der Leitlinien (68) Erziehungswissenschaft und Psychologie (63) Überblick über SMT-A (48) Fachbereich Erziehungswissenschaft und (47) für komplexe Zahlen (46) to first slide (46) Back to first (46) Klasse für komplexe (46) View graphic version (45) zurück zur Übersicht (44) Überblick über SMT (44) Zurück zur Startseite (36) HTML-Dateien selbst erstellen (36) Freie Universität Berlin (36) Der rote Faden (35) Garten und Botanisches (33) und Botanisches Museum (33) Index Leitlinien der (32) Botanischer Garten und (32) Kurzinfo mit Signatur (32) Botanisches Museum Berlin-Dahlem (32) Zurück zum Inhaltsverzeichnis (31) zur Übersicht über (30) Zum Starten hier (29) Starten hier klicken (29) Technische Universität Chemnitz (28) Übersicht über Kapitel (28) Übertragungstechnik und Bitübertragungsschicht (27) of Industrial Relations (27)

Tabelle 3: Die je 40 häufigsten in Hyperlink-Ankern vorkommenden Trigramme (Abfolgen von je drei Token) aus S1 – S3

Navigationshilfen sie dem Benutzer geben. Amitay (2000) untersucht diesen Aspekt ebenfalls, bezieht sich jedoch auf Bigramme. In ihren Korpora findet sie unter den 20 häufigsten, in Linkankern enthaltenen Bigrammen u. a.: „home page“, „return to“, „back to“, „more info“, „click here“ und „to the“.

Tabelle 3 zeigt die jeweils 40 häufigsten Trigramme aus S1 – S3. Diese enthalten ebenfalls räumliche Deiktika (speziell hierzu de Saint-Georges, 1998): „Zurück zum Anfang“, „Zurück zur Homepage“, „Zurück nach oben“. Auffällig sind die insgesamt 31 Vorkommen von „zurück“ in den 120 Trigrammen. Bei genauerer Betrachtung erscheint diese Häufung von Trigrammen, die zurück zu einer Übersichts- oder Indexseite führen, logisch: Neben Übersichtsseiten existieren – betrachtet auf einer sehr abstrakten Ebene – Seiten, die Inhalt vermitteln. Der Rücksprung von einer solchen Inhaltsseite – eine zentrale Handlung bei der Web-Navigation – zur Übersicht wird metaphorisch durch das Zurückspringen oder Zurückgehen zur Übersicht ausgedrückt. Betrachtet man die einzelnen Vorkommen von Trigrammen, die „zurück“ enthalten, werden die Übersichtsseiten mit einem relativ kanonischem Vokabular bezeichnet: „Leitseite“, „Homepage“, „Anfang“, „Startseite“, „Übersicht“, „Seitenanfang“, „Inhaltsverzeichnis“, „Hauptseite“ etc. Die meisten anderen Trigramme entstammen aus Linkankern, die nach vorne verweisen, also von einer Übersichtsseite zum Inhalt führen, z. B. „Traditionelle Chinesische Medizin“, „Absatztypen und Textgestaltung“, „[Zum] Starten hier klicken“.[17] Bei letzterem Trigramm fällt auf, dass es das Einzige ist, das „klicken“ enthält. Der in Ratgebern für gute Webgestaltung häufig gescholtene Imperativ „klicken Sie hier“ oder „hier klicken“ taucht de facto also nicht (mehr) häufig auf (S1: „Hier klicken“, Pos. 88 in den Bigramm-Häufigkeiten mit 101 Vorkommen, S2: kein Vorkommen in Bigrammen oder Trigrammen, S3: „hier klicken“, Pos. 101 in den Trigramm-Häufigkeiten mit 37 Vorkommen).

4.4 Merkmale konzeptioneller Mündlichkeit

In Abschnitt 2 werden verschiedene Merkmale aufgezählt, die in der „klassischen“ CMC-Literatur, die sich vornehmlich mit E-Mails, Newsgruppen und dem Internet Relay Chat beschäftigt, als Kennzeichen für konzeptionelle Mündlichkeit dargestellt werden. Die Vorkommen derjenigen Merkmale, die sich maschinell erkennen lassen, werden in den folgenden Abschnitten analysiert.

4.4.1 Smileys

Das erste Merkmal, das zweifelsfrei eine gewisse Nähe zur konzeptionellen Mündlichkeit evoziert und sehr häufig in der Chat-Kommunikation (vgl. etwa Schlobinski, 2000) genutzt wird, sind Smileys. Zur automatischen Erkennung von Smileys wurde die von James Marshall[18] zusammengestellte, mit 2 100 Einträgen vermutlich weltweit umfangreichste Smiley-Liste, in das Analysesystem integriert. Etwa 160 in dieser Liste enthaltene Smileys wurden nicht berücksichtigt, da sie Zeichen bzw. Zeichenketten darstellen, die nicht eindeutig als Smileys zu identifizieren sind.[19]

---

[17] „Zum Starten hier klicken“ ist die in Microsoft Powerpoint verwendete Floskel, um in das HTML-Format konvertierte Powerpoint-Folien zu starten. 1 706 der in S3 enthaltenen Dokumente (S1: 209 Dokumente, S2: 50 Dokumente) enthalten im Kopf der jeweiligen HTML-Datei einen Verweis (z. B. `<meta name="GENERATOR"content="Microsoft PowerPoint 9">`), dass diese Datei von Microsoft Powerpoint erstellt wurde. Auf MS Powerpoint gehen auch die in S3 sehr häufigen Trigramme „Zurück zur ersten“ und „zur ersten Folie“ zurück.

Anhand einer empirischen Analyse dieser Meta-Elemente und der assoziierten Dokumente könnte eine in Storrer (2001b) aufgestellte These verifiziert werden: „Viele Entwicklungsumgebungen liefern vorgefertigte Schablonen zur Homepage-Gestaltung mit, die nur noch aufgefüllt werden müssen. Diese tragen, da sie ohne großen Zeit- und Geldaufwand zur eigenen Homepage führen, natürlich erheblich zur Herausbildung und Verfestigung von Strukturierungs- und Gestaltungsmustern im WWW bei.“

[18] Erhältlich unter `http://www.astro.umd.edu/~marshall/`.

[19] Beispielsweise `0`, das in o. a. Liste als der aus der Science Fiction-Reihe „Star Wars“ bekannte Todesstern geführt wird und auch eher in den Bereich der (minimalistischen) ASCII-Art gehört. Siehe zur Verwendung von ASCII-Art in Chatsystemen beispielsweise Haase et al. (1997, S. 78); im World Wide Web kommt ASCII-Art als Stilmittel m. E. nur sehr selten vor,

S1	;-) (412) :-) (345) ;) (85) :) (79) :-)) (54) :-((37) ;-)) (33) ;)) (21) :)) (17) :o) (13) :((13) ;o) (11) :-))) (9) =:-) (8) :))) (7) -:-) (7) :-(((6) =) (5) (c: (5) ;:-) (5) =:-((5) :-D (4) |o| (4) ;-((4) :o((4) 8-) (4) (-: (3) [* (3) (: (3) :-> (3) ;((3) :- (3) :-o (3) :-X (3) :-)))) (3) ;-))) (3) :-P (3) {(:-) (3) :D (2) ;-} (2) (|) (2) :'-((2) C= (2) 8-| (2) :-@ (2) [*] (2) :=) (2) :-() (2) :-{) (2) :-{} (2) :-] (2) :-[(2) ;)))))) (2) :-| (2) >:-> (2) >;-> (2) [:-) (2) :-Q (2) (@ (2) {:-) (2) =0) (2) (:-) (2) :-, (2) :-/ (2) --. (1) :*) (1) :,((1) :-# (1) :-& (1) :-O (1) :-7 (1) :-C (1) :-O (1) :O) (1) :^D (1) :~) (1) :-x (1) >:) (1) O-) (1) :'-) (1) :-p (1) X-((1) [#] (1) "< (1) [:] (1) '! (1) :* (1) :-e (1) :-)8 (1) :-)= (1) :-)~ (1) :< (1) :> (1) ^-^ (1) :-~) (1) :-|| (1) c[] (1) ;))) (1) {*} (1) |-) (1) |-| (1) :[(1) :-} (1) :)))) (1) <3 (1) :| (1) *<:-) (1) <:-I (1) :} (1) :-W (1) >:-((1) <|-) (1) (^-^) (1) #-) (1) %-) (1) %-6 (1) :-)-8 (1) %-| (1) :-)... (1) :-{)} (1) B:-) (1) %-} (1) 8:-) (1) '-) (1))8-) (1) *:o) (1) +:-) (1) (:I (1) :-)))))) (1) +-:-) (1) ;-)))))) (1) ,-) (1)
S2	:-) (111) ;-) (63) :) (25) :-((17) ;) (16) :-)) (11) ;-)) (7) :((5) :o) (5) ;-))) (4) -((3) ._) (2) :-o (2) :// (2) :-x (2) :<= (2) :-| (2) <:>== (2) >-^);> (2) *<:-) (2) %*@:-((2) ;-D (1) d:-) (1) (-: (1) ;))) (1) :] (1) :)))) (1) 8*) (1) ;o) (1) ;)))) (1) H-) (1) (: (1)
S3	:-) (37) ;-) (26) :) (7) :-{ (4) ;) (4) :-((3) :-)) (2) :Q (1) %-) (1) ;-} (1) :^) (1) :o) (1) =0) (1) H:) (1) :-(((1) :((1) :-)))))))) (1)

Tabelle 4: In S1 – S3 enthaltene Smileys und ihre jeweiligen Vorkommen

Die Vorkommen von Smileys in den drei Stichproben sind sehr heterogen. Auf den studentischen Homepages existieren insgesamt 1 353 Smileys in 806 Dokumenten (Maximum: 61), auf den Homepages der Universitätsmitarbeiter insgesamt 298 (in 178 Dokumenten, Maximum: 16), in den tiefen Dokumenten 93 Smileys in 58 Dokumenten (Maximum: 6), vgl. Tabelle 4. Die deutlich häufigeren Vorkommen in S1 unterstützen die bei der Betrachtung der Wortfrequenzen aufgestellte These, dass die in dieser Stichprobe vorhandenen Dokumente einen eher informellen, konzeptionell mündlichen Stil aufweisen.

Im Folgenden betrachte ich den Inhalt derjenigen Dokumente, die pro Stichprobe die jeweils häufigsten Vorkommen von Smileys enthalten. S1: An erster Stelle befindet sich eine Smiley-Liste, d. h. ein Dokument, das unterschiedliche Varianten von Smileys erläutert (61 Vorkommen), gefolgt von einer weiteren derartigen Liste (51 Vorkommen), ein konzeptionell sehr mündlicher, 1 635 Wörter umfassender Bericht („DRM-Alpentour '99 – die 1. Woche“, `http://www.kawo2.rwth-aachen.de/~bossi/at99/bericht1.html`) über einen Motorradausflug in die Alpen (20 Vorkommen), eine etwa 70 Einträge umfassende Seite eines Gästebuchs (16 Vorkommen) und ein Bericht („Lebenslauf eines Suechtigen“, `http://wwwstud.uni-giessen.de/~s399/sucht3.html`) über die persönliche Erfahrung mit Computern. S2: eine umfangreiche Aufstellung von Computer-bezogenen Anekdoten („Dümmste anzunehmende User – Die Sammlung“, `http://www.uni-giessen.de/~gc1091/jokes/dau.htm`, 16 Vorkommen), eine Smiley-Liste (13 Vorkommen), die persönliche Homepage eines wissenschaftlichen Mitarbeiters im Bereich anorganische Chemie (8 Vorkommen), ein Kapitel einer soziologischen Zwischenprüfungsarbeit („Kommunikation und Rollenverhalten in Virtuellen Realitäten anhand des MUDs Xyllomer“, `http://www.uni-giessen.de/~g31048/zpa/untersuchung.html`), in dem 6 Vorkommen von Smileys in Zitaten von Benutzern des untersuchten MUDs enthalten sind und ein „Kurzer Vietnam Reiseführer“ (`http://www.iwr.uni-heidelberg.de/~mennicke/`

etwa in der Überschrift „_oO(Ein kleines Werk)Oo._“, `http://www.tu-chemnitz.de/~kirst/`.

viet/, 5 246 Wörter) mit 4 Vorkommen. S3: Eine im HTML-Archiv einer Mailing-Liste, die sich mit dem Textsatzprogramm TEX beschäftigt, vorhandene, sehr lange E-Mail (6 Vorkommen), die HTML-Version der Diplomarbeit einer Psychologin („Kommunikationsstrukturen und Persönlichkeitsaspekte bei MUD-Nutzern“, http://www.tu-chemnitz.de/phil/psych/professuren/sozpsy/Mitarbeiter/Utz/Diplom1.htm, 5 Vorkommen), zwei weitere E-Mails aus dem Archiv o. g. Mailing-Liste (5 bzw. 4 Vorkommen) und ein „Kommentar zur Evaluation des E-Mail-Projektes Sommersemester 1998“ (http://www.tu-dresden.de/sulifg/daf/mailproj/komenta1.htm, 4 Vorkommen).

Hier werden zwei unterschiedliche Aspekte[20] deutlich: Einerseits enthalten Dokumente, die auf den persönlichen Homepages von Studierenden angeboten werden (z. B. der Bericht über die Reise in die Alpen oder die humoristische Darstellung der eigenen Computersucht), sehr viele Smileys. Längere Texte werden mit diesem Stilmittel ausgestattet, um den saloppen und nicht zu ernst gemeinten Stil dieser Beiträge zu unterstreichen. Der zweite Aspekt ist die Tatsache, dass selbst in Dokumenten, die von Universitätsangehörigen, meist wissenschaftlichen Mitarbeitern, angeboten werden, Smileys enthalten sind. Wenn es um einen Themenbereich wie Internet-Kommunikation (die beiden o. g. wissenschaftlichen Arbeiten zur MUD-Kommunikation) geht, tauchen Smileys zwangsläufig in Beiträgen oder Zitaten der MUD- oder IRC-Benutzer auf, jedoch enthält die in S3 enthaltene Evaluation des E-Mail-Projekts vier Smileys, die vom Projektleiter selbst eingefügt worden sind. Das Dokument besteht aus einer Auflistung von Anmerkungen, die aus Evaluationsbögen zu einem E-Mail-Projekt im Bereich „Deutsch als Fremdsprache“ stammen. Der Leiter hat die Anmerkungen der Studierenden kommentiert, und er beantwortet den informellen Stil der Fragen ebenso informell, was durch den Einsatz von Smileys noch verstärkt wird. Meiner Einschätzung nach ist dies ein Anzeichen dafür, dass Smileys als Ausdruck der Ironie oder um eine humorvolle Bemerkung zu unterstreichen in Zukunft auch in eher offiziellen Webseiten (vs. privaten, studentischen Homepages) häufiger auftreten werden.

4.4.2 Iterationen

Tabelle 5 enthält die jeweiligen – oftmals sehr stark an die Comic-Sprache[21] angelehnten – Vorkommen von Iterationen. Dies sind Reduplikationen einzelner Buchstaben eines Wortes (z. B. „soooo“, „Suuuper“, „Tschüßiiiii“) zum Zwecke der Emulierung von Prosodie, insbesondere der Sprechgeschwindigkeit (vgl. Haase et al., 1997, S. 67 f.). Wie bei den Smileys enthält S1 mit 834 Iterationen in 435 Dokumenten die mit Abstand meisten Vorkommen. S2 enthält 84 Iterationen in 55 HTML-Dateien und S3 lediglich 20 Vorkommen in 15 Dokumenten.

Betrachten wir auch hier die jeweiligen in S1 – S3 enthaltenen Dokumente mit den häufigsten Iterationen. S1: Ein äußerst umfangreiches (28 935 Wörter) Textkorpus von Plattenkritiken[22] (26 Vorkommen), ein Dokument mit Texten und zugehörigen Gitarrenakkorden der Gruppe „Tocotronic“ (15 Vorkommen), die ins Deutsche übersetzten, in der Internet- und UNIX-Szene sehr bekannten Geschichten des „Bastard Operator from Hell“ (http://www.wohnheim.uni-ulm.de/~amaenn/bofh.htm, 15 Vorkommen), das dynamisch generierte Logfile eines Web-basierten Chatraums (15 Vorkommen) und „Die einzig offiziöse Spezifikation für BTML (und ehemals JTML)“[23] (http://www.wohnheim.uni-ulm.de/~schabi/usenet/btml.jtml.html, 14 Vorkommen). S2: Die bereits

[20] Die Feststellung, dass Smiley-Listen viele Smileys enthalten, ist trivial. Jedoch sind durch diese Listen die Vorkommen eher unüblicher Smileys zu erklären, die in Tabelle 4 aufgeführt sind. Auf die Vorkommen von Merkmalen konzeptioneller Mündlichkeit in Gästebüchern gehe ich an dieser Stelle nicht ein und verweise stattdessen auf Diekmannshenke (2000).

[21] Siehe hierzu auch die IRC-Protokolle in Sassen (2000) und Schmidt (2000, S. 123).

[22] Zu dem Artikel „Die Plattenkritik: eine empirische Textsortenanalyse“ (1999), Jannis Androutsopoulos. In: Jens Neumann (Hrsg.). *Fanzines II. Wissenschaftliche Betrachtungen zur Medienlandschaft der Subkulturen*. Mainz: Ventil Verlag.

[23] BTML und JTML stehen für „Bizarre“ bzw. „Joke Talk Markup Language“. Dies sind Auflistungen von XML- bzw. SGML-ähnlichen Tags, die erstmals im Usenet – im deutschsprachigen Bereich in den Gruppen de.talk.bizarre und de.talk.jokes – aufgetaucht sind. Einige Elemente werden sogar mit einer korrekten XML-Syntax definiert, wie beispielsweise <beschwichtigung>, das die Attribute ups, wieKonnteIchNur oder immerPassiertMirDas besitzen darf.

S1	sooo (34) Hmmm (17) soooo (11) hmmm (11) aaaaa (11) Gääääähn (10) gaaaanz (9) gaaanz (9) Mmmm (9) Ahhh (8) Naaah (8) gaaaaanz (6) hmmmm (6) aaaaaa (6) grrr (6) seeeehr (6) Fickööön (5) naaaaaaaaaaaaaaaaaaaaaaaaaaa (5) ööööööööööööööööööö (5) pssst (5) Sooo (5) Aahhh (4) Ääääätsch (4) Pffft (4) Auuuu (4) Tadudeldidudeldidudeldiduuuuu (4) BOOOM (4) schööön (4) cooool (4) blahaaaa (4) Hmmmm (4) Pfüüüüüp (4) Alteeer (4) viiiel (4) Waaaas (3) Seeesel (3) Uuuups (3) soooooo (3) SOOO (3) Ahhhh (3) öööööööööööööööööööööööööööööööööööööö (3) Prfffz (3) Jaaaaa (3) Ohhh (3) neee (3) BWLIII (3) Ooooh (3) Pssst (3) uuuu (2) Gaaanz (2) Ooohhh (2) Äquatorrr (2) noooch (2) Gäääähn (2) hmmmmmmmm (2) Halllihallllllllllllllllloooooooooooooooo (2) grrrrr (2) ahhhhh (2) gaaaaaaanz (2) ohhh (2) Sorrrry (2) AjrRawkFGwwAAAAACgkQQTauo (2) räääähh (2) brrrrrilliant (2) aaaaaaaa (2) ähhh (2) WAAAAHNSSINNS (2) Aaahhh (2) Wuuduuu (2) URZsssssssthhhh (2) Ahhhahhhahhh (2) Määäh-Synthesizer (2) hiiiiiieer (2) HMMM (2) Ahhohhhh (2) SCCCCCCCCCCCCCHMAAAAAAAAAATTZZ (2) grrrrreat (2) groooooßen (2) lieeebe (2) seeehr (2) aaaa (2) Pieeep (2) tatatataaa (2) sooooo (2) Snnnniggers (2) Jaaa (2) Tüttüüüt (2) Fuuuunnnnnnn (2) coool (2) rrrrrrrrroor (2) Mmmmh (2) herrrliche (2) Pfffft (2) rooroororrorruuuuum (2) JUBELLLL (2) aaaaaaaaaaa (2) GAGNEEEE (2) vieeel (2) aaaaff (2) aaaaaaaaaa (2) FZZZZT (2) AAAARRGGHH (2) Naaa (2) nööö (2) GAAANZ (2) Aaaah (2) Chhrrrr (2) Mööööööööööp (2) muuuussss (2) youuuuuuuuu (2) wunderschööön (2) Hmmmmm (2) psssst (2) firrrst-a-place (2) herrrlich (2) vieeeel (2) Mööööönsch (2) Ähhh (2) Allsssooo (2) meeeeeeeeeeehr (2)
S2	sooo (4) gaaanz (4) gaaaaaaanz (4) Mussssssss (3) Mussssssss (3) Klaaasse (2) daaaaa (2) Hoschiiiii (2) Jooooooooo (2) Suuuper (2) uuuuh (2) blubbblubb (2) Hardmannnnshennnnnnnnnnnnnn (2) gaaaaanz (2) Aaaaaaaaaabstand (2) Aaaaaaaaaaaaaaaaaaaaaaaaaabstand (2) AHV-xxxx (2) tiiieerisch (2) AAAaaaarrrghgghghgh (1) uuuuuuhhhhhhhhhooaaahhhhh (1) Aaaaaaaaaahhhhhhhhhhhhh (1) Ahhhh (1) brennnt (1) Haaa (1) schaaaaa (1) uuuuuiiiiIIIIIII (1) AAAAT (1) natüüürlich (1) sooooo (1) gaaasanz (1) Tschüßiiiiii (1) Jaaa (1) Hoschiiiiiii (1) Hoschiiiiii (1) gaaaaannz (1) süüüüüß (1) SOORRRYYY (1) seeeeeeehr (1) hmmm (1) hiiiiiiiilfe (1) wiesooooo (1) gaaaanz (1) vistaaa (1) loginnnnnnnnnn (1) WAAAS (1) Soooo (1) Uffff (1) Gaaaaz (1) Tierraaa (1) AIII (1) Hoschiiiiii (1) Reeeeegen (1) Sooooo (1) Duuuh (1) Hmmmm (1) AAAAEH (1) Brrr (1) Giiiiiiiiiiiiiießen (1)
S3	Hmmm (4) RRRrrrr (2) huuuderts (1) viiieele (1) rassschen (1) GROOOOOSS (1) ZUUU (1) Jaaaaaaaaaaaa (1) tiriliii (1) AAAaaarrrrggghhhhhhhh (1) AAAAAAAAAAAATTTTTTTTTTT (1) mmmmmh (1) yäääääääääh (1) BRRR (1) TSCHÜÜÜÜÜÜÜÜSS (1) Duuuuuuu (1)

Tabelle 5: In S1 – S3 enthaltene Iterationen und ihre jeweiligen Vorkommen (die nur jeweils einmal existenten Vorkommen aus S1 sind aus Platzgründen nicht enthalten)

im Abschnitt über Smileys erwähnte Anekdotensammlung „Dümmste anzunehmende User - Die Sammlung“ (13 Vorkommen), eine Webseite mit Bildern des Labors eines wissenschaftlichen Mitarbeiters aus dem Bereich der angewandten Physik (3 Vorkommen, u. a. „Das Platzangebot ist so groß, daß ich mir mit Almut, Kai und Michael (Hoschiiiiiii) das Büro teilen darf – ätsch!“), ein Dokument, das lediglich ein Schwarz-Weiß Photo einer Bar sowie die Satzteile „Die Rehlein beten zur Nacht,“ (oberhalb des Bildes) und „hab acht!“ (unterhalb) zeigt (`http://www.tu-harburg.de/~w3em/Rehlein/gebet7.html`, 2 Vorkommen[24]), eine weitere Bildergalerie des o. g. Physiklabors (2 Vorkommen) und eine Link-Sammlung zum Thema Linux (2 Vorkommen, u. a. „Tja das ist ein absolutes Mussssssss!“). S3: Eine direkt aus der Microsoft Word Quelle nach HTML konvertierte, etwa 60 Papierseiten (14 647 Wörter) umfassende Broschüre für Anfänger des Studiengangs Bauingeneurwesens an der TU Dresden (4 Vorkommen: zweimal die Kapitelüberschrift „RRRrrr“ sowie „Hmmm“ an zwei Satzenden), eine Informationsseite des „Auslands-AK“ der Fachschaft Jura an der Universität Heidelberg (2 Vorkommen, „Jaaaaaaaaaaaa, die Fachschaft hat dieses Jahr auch einen “Ausland-AK”!!!!!! Was sich hinter diesem schwammigen Begriff versteckt (das Ausland ist bekanntlich GROOOOOSS), soll hier mal geklärt werden.“), ein zehn Einträge umfassendes Gästebuch und eine Bibliographie zum Thema Napoleon (ein Vorkommen, das auf einen Tippfehler zurückzuführen ist: „Deutschland in der Weltpolitik des 19. und 20. Jahrhuuuderts“).

Neben Smileys scheinen auch Iterationen vornehmlich in den persönlichen Homepages von Studierenden vorzukommen (835 Vorkommen vs. 84 bzw. 20). Hierbei fällt auf, dass viele der in S1 enthaltenen Dokumente, die mehrere Iterationen beinhalten, nicht primär für das World Wide Web geschrieben wurden: Das Textkorpus mit Plattenkritiken basiert auf Fan-Magazinen der untersuchten Musikstile, und Dateien mit Texten und Gitarrenakkorden von Musikgruppen werden schon seit den achtziger Jahren in den Newsgruppen des Usenet getauscht; aus dieser Zeit stammen auch die amüsanten Geschichten des „Bastard Operator from Hell“. Durch eine derartige Alternativverwertung bzw. -veröffentlichung werden also unweigerlich sprachliche Phänomene in das World Wide Web transportiert. Dieser Prozess tritt besonders deutlich bei der „Spezifikation für BTML“ hervor: Ohne den Erfolg des World Wide Web wäre XML nicht entwickelt worden, so dass ein derartiger humoristischer Einsatz von XML-Elementen, denen ad hoc Namen zugewiesen werden, die den entsprechenden Kontext oder die Aussage einer Nachricht unterstreichen, undenkbar gewesen wäre (weitere Beispiele: `<bedauernd_die_schultern_hochzieh>` oder `<SCCCCCCCCCCCCHMAAAAAAAAAAATTZZ>`: „Ein dicker Kuß. Wird vom “Opfer” oft mit `<Ohrfeige>` wieder geschlossen.“). Die technischen Details des Mediums provozieren also durchaus neue und kreative sprachliche Ausdrucksformen.

Die wenigen Vorkommen von Iterationen in S2 (hier etwa die o. g. Unterschrift der Photos des Physiklabors) und S3 scheinen Ausnahmen darzustellen. Meiner Ansicht nach wird man in den kommenden Jahren auf derartigen Dokumenten, die eher offizielle Informationen einer Organisation als persönliche Daten darstellen, immer weniger konzeptionell mündliche Merkmale finden (Smileys könnten hier evtl. eine Ausnahme darstellen, vgl. Abschnitt 4.4.1), da einerseits die Glaubwürdigkeit einer Quelle hierunter leider und andererseits der Aspekt des Neuen und Spannenden sich schon bald abgenutzt haben könnte.

4.4.3 Emphasen

Isolierte Verbstämme, Iterationen und andere sprachliche Spezifika tauchen in der Internet-Kommunikation häufig in einer markierten Form auf, die ich generisch als Emphase bezeichne. Die Markierung erfolgt dabei durch die Einschließung eines Wortes (oder einer Sequenz von Worten) in

[24] Das zur Analyse der Samples eingesetzte Perl-Skript erkannte in diesem Dokument die beiden Iterationen „Aaaaaaaaaabstand“ sowie „Aaaaaaaaaaaaaaaaaaaaaaaaaaaaabstand“, die – in der gleichen Farbe wie der Hintergrund gehalten – vom Autor als Blindtext eingesetzt wurden, um die beiden Satzbestandteile links- bzw. rechtbündig anzuordnen.

S1	*FERNSEHEN* (14) *grins* (10) *hihi* (8) *HÖRFUNK* (8) *verbeug* (6) *die* (3) *gg* (3) *seufz* (3) _nicht_ (3) *PLONK* (3) *lol* (3) /Rechtsberatung/ (3) *Der* (2) *Patsch* (2) /SGML/ (2) *smile* (2) *frechgrins* (2) *bg* (2) *fg* (2) _alle_ (2) *so* (2) *SABBER* (2) *DER* (2) *sehr* (2) *PATSCH* (2) /10/ (2) *ggg* (2) *knuddel* (2) _sprache_ (2) *wink* (2) *stutz* (2) *drei* (1) *FUSSBALL* (1) /talktalktalk/ (1) *Symmetrieachse* (1) *liebe* (1) _snabel_ (1) *dickesLob* (1) *etwas* (1) *schmaatz* (1) /reise/ (1) *fühlemichgebauchpinselt* (1) *gggg* (1) *Gähn* (1) *verkaufen* (1) *aufgeregt* (1) /laxt/ (1) /issues/ (1) /eng/ (1) *bell* (1) *heul* (1) *schmunzel* (1) *06* (1) *are* (1) *räckel* (1) *etwa* (1) *kein* (1) *bis* (1) *GGG* (1) *schleck* (1) *schluck* (1) *trommel* (1) *alle* (1) *grübel* (1) *inSicherheitfuehl* (1) *KEUCH* (1) /Kleinhirnblutungen/ (1) *freu* (1) _underlined_ (1) *grinst* (1) *ätsch* (1) /selfhtml/ (1) *giggel* (1) *BLINDMAIL* (1) *tnx* (1) *hexhex* (1) *Lecker* (1) *YES* (1) /11/ (1) *springinLuft* (1) *ruhe* (1) *ubs* (1) *flame* (1) *winkewinke* (1) /ei/ (1) *SUPERSKIPAUSCHALE* (1) *nicht* (1) *freudig* (1) /wg/ (1) *keinen* (1) *PRIVAT* (1) *hoff* (1) /ops/ (1) /winamp/ (1) *GreatIsTheLord* (1) /flyer/ (1) _NICHT_ (1) /darkzone/ (1) *übs* (1) *lächergaaaaaaanzbreit* (1) _null_ (1) *neu* (1) /ski/ (1)
S2	*Andy* (5) *andy* (3) *grins* (2) _nicht_ (2) _eingeschränkt_ (2) /Kurs/ (2) *such* (2) *gaaaaannz* (1) *gacker* (1) *difference* (1) *besten* (1) *nicht* (1) _einige_ (1) *Beraterin* (1) *grummel* (1) /vor/ (1) /neues/ (1) *Uffff* (1) /Inhalt/ (1) *funktionierende* (1) *aufgeregt* (1)
S3	*kein* (1) *lach* (1) *lall* (1) _impulsiv_ (1) *kostenlos* (1) *zu* (1) *sollte* (1) _japanisches_ (1) _first_ (1) _ohne_ (1) _ganze_ (1) *erforderliche* (1) /agrep/ (1) /95/ (1) /aU/ (1) *bindend* (1) *ohne* (1) *jammer* (1) *vollständig* (1) _neue_ (1) *lol* (1) /dev/ (1) *bitte* (1) *not* (1)

Tabelle 6: In S1 – S3 enthaltene Emphasen und ihre jeweiligen Vorkommen

Sonderzeichen; vornehmlich wird der Asterisk, `*`, eingesetzt, doch der Unterstrich, `_`, und der Slash, `/`, sind ebenfalls gebräuchlich (siehe hierzu u. a. Haase et al., 1997, S. 67 ff.).

Tabelle 6 zeigt die Vorkommen von Emphasen in den drei Stichproben. S1 liegt auch hier mit 182 Vorkommen in 92 verschiedenen Dokumenten an der Spitze, wohingegen S2 mit 38 Emphasen in 16 HTML-Dateien und S3, 24 Vorkommen in 14 Dokumenten, je weniger als ein Drittel dieser speziell markierten Wörter enthalten. Auch hier betrachten wir die in den Stichproben enthaltenen Dokumente mit den meisten Vorkommen. S1: Mit 23 Vorkommen an der ersten Position ist ein Dokument, das aus einer großen Anzahl von Videotextseiten[25] besteht, die mit Hilfe einer speziellen TV-Karte und entsprechender Software nach HTML konvertiert wurden, an zweiter Stelle liegt die bereits erwähnte „offiziöse Spezifikation für BTML“ (7 Vorkommen), gefolgt von einer tagebuchartigen studentischen Homepage (7 Vorkommen, z. B. „*schleck* Grad gab's lecker Gries mit selbstgekochtem Apfelmus (Julie hatte die Äpfel eigentlich schon weggeschmissen, aber als Mus sind sie mir lieber als als Müll) und Zimt.“), gefolgt von einem weiteren Gästebuch mit 40 Einträgen (6 Vorkommen, beispielsweise „deswegen hatte ich keine wahl, ich meine bei der vergabe des awards *grins* äh, nochmal *grins* top design, doppelt fun !!“). S2: 17 der 32 in S2 enthaltenen Emphasen gehen auf den bereits mehrfach angesprochenen Text „Dümmste anzunehmende User – Die Sammlung“ zurück, darauf folgt – mit jeweils einem Vorkommen – eine Anleitung, wie an der Universität Gießen unter Linux eine („*funktionierende*“) PPP-Verbindung mit dem Rechenzentrum herge-

[25] Da die Möglichkeiten der typographischen Auszeichnung auch im Teletext sehr begrenzt sind (es existiert kein Fettdruck, lediglich eine geringe Anzahl Farben stehen zur Verfügung), werden offenbar auch in diesem Medium Asteriske eingesetzt, um wichtige Wörter zu markieren, beispielsweise `*FERNSEHEN*`.

stellt werden kann, eine weitere Webseite stellt eine Art „Fragen Sie den Experten“-Service[26] dar, der von einem an der Universität Augsburg ansässigen Professor angeboten wird. S3: An erster Stelle steht hier die HTML-Version einer FAQ-Liste der Newsgruppe `bln.markt` (5 Vorkommen), gefolgt von einer im Web-Archiv einer Mailing-Liste enthaltenen E-Mail (4 Vorkommen) sowie einem Gästebuch (ebenfalls 4 Vorkommen).

Emphasen treten nicht annähernd so häufig auf wie Iterationen, doch ist die Bandbreite dieses Stilmittels relativ groß: In den für S1 in Tabelle 6 notierten Vorkommen existieren isolierte Verbstämme („*grins*“, „*hoff*“, „*flame*“), spezielle Abkürzungen (ausführlich zu Akronymen in Bezug auf E-Mail siehe Ziegler, 2001) oder Slang-Ausdrücke („*PLONK*“, „*bg*“, „*gg*“, „*tnx*“), eine dialektale Variation („_snabel_“), tw. iterierte Wortsequenzen („*lächergaaaaaaanzbreit*“, „*fühlemichgebauchpinselt*“) oder vollständige Großschreibung, die häufig als emuliertes Schreien interpretiert wird („*KEUCH*“). Darüber hinaus resultieren einige der gefundenen Iterationen aus einer nicht vollständigen Übertragung des jeweiligen Dokuments in die Hypertext Markup Language. Die `bln.markt` FAQ-Liste (`http://www.math.fu-berlin.de/user/guckes/faq/bln.markt/`) macht durchaus Gebrauch von der technischen Möglichkeit, Überschriften als solche auszuzeichnen und Fettdruck zu benutzen, doch eine Emphase wie in „... sondern den Artikel erneut *vollständig* (mit Korrektur) posten.“ bleibt verbatim erhalten, obwohl ein kursiver Schriftschnitt aus typographischer und gestalterischer Sicht angebrachter wäre.

4.4.4 Isolierte Verbstämme

S1	grins (5) grab (2) laber (2) quietsch (2) würg (2) stöhn (2) schnief (2) seufz (2) sniff (1) röchel (1) nörgel (1) hust (1) schmatz (1) knuddel (1)
S2	– / –
S3	– / –

Tabelle 7: In S1 isoliert enthaltene Verbstämme und ihre jeweiligen Vorkommen

Im vergangenen Abschnitt wurden bereits isoliert auftauchende Verbstämme angesprochen, die vor allem in der Chat-Kommunikation häufig eingesetzt werden, um auf knappe Weise Zustände oder Gefühlsregungen des Autors bzw. Benutzers auszudrücken (vgl. u. a. Haase et al., 1997, S. 65). Speziell markierte Zeichenketten – o. a. Emphasen – können automatisch mit Hilfe regulärer Ausdrücke erkannt werden. Um auch die nicht markierten Vorkommen isolierter Verbstämme ausfindig zu machen, wurde eine Liste von 5 438 Verbstämmen in den Erkenner integriert und die jeweiligen Treffer manuell verifiziert. Die wenigen Vorkommen (vgl. Tabelle 7) isolierter Verbstämme befinden sich in der Stichprobe, die persönliche Homepages Studierender enthält und können fast ausnahmslos dem Bereich der Comic-Sprache zugeordnet werden. Zwei der insgesamt 25 Vorkommen stammen aus einer Webseite (`http://www.rzuser.uni-heidelberg.de/~mdemiral/grue%DFe.htm`), die offenbar eine Grußbotschaft[27] an die Schwester des Autors ist:

[26] Auf `http://www.physik.uni-augsburg.de/~ferdi/fragen/frage3.html` wird gefragt, weshalb Menschen in großer Höhe schneller bzw. langsamer altern. Die deutschsprachige Antwort enthält einen englischsprachigen, per Copy & Paste eingefügten Text „aus dem Internet“, in dem der Beleg „*difference*“ enthalten ist.

[27] Ein Photo des aus der Science Fiction Reihe „Star Wars“ bekannten Jedi-Lehrmeisters Yoda ziert den Hintergrund dieser Webseite. Obwohl er bereits mehr als 900 Jahre alt ist, beherrscht Yoda die Syntax des Deutschen (bzw. ursprünglich Englischen) nicht, was sicherlich die Motivation für den syntaktischen Aufbau von „gesagt hat er das ‚das er dich liebt“

Hu hu
abla das ist fuer dich Lustig was?
das kann man ganz schnell machen
ausserdem ist es garnicht so schwer
Hey :gesagt hat er das ,das er dich liebt
schmatz und knuddel
liebe schwester

Zwei weitere Vorkommen befinden sich in einem fiktiven Dialog (`http://wwwstud.uni-giessen.de/~s6473/intelligenz.html`), der die Schwächen automatischer Dialogsysteme verdeutlichen soll. Nach einigen verzweifelten Versuchen endet der nur schleppend stattfindende Dialog mit „Anrufer: .... stöhn ... röchel ....!“. In der Bildergallerie eines Strandurlaubs wird ein Photo, das den Bau einer Sandburg zeigt, unterschrieben mit „grab grab ...“.

Nur auf etwa der Hälfte der Dokumente, die derartige isoliert auftretende Verbstämme enthalten, sind auch Hinweise darauf zu finden, dass die jeweiligen Autoren häufiger Chat-Systeme (IRC oder Web-Chats, zu letzteren etwa Storrer, 2001a) benutzen. Dies könnte einerseits die Vermutung nahelegen, dass sich das Phänomen der isolierten Verbstämme derzeit immer mehr – auch unter den Nicht-Chattern – verbreitet, oder aber, dass es sich lediglich um ein sehr seltenes sprachliches Phänomen handelt, das eher zufällig bei der oftmals sehr raschen Produktion der betroffenen Seiten entstanden ist. Eine mit einigen Jahren Abstand durchgeführte vergleichende Untersuchung könnte evtl. eine Antwort auf diese Frage liefern.

4.4.5 Slangausdrücke

S1	linx (21) dau (20) imho (7) rpg (6) motd (6) bofh (5) rl (5) btw (3) lol (3) snafu (2) ppl (2) cfv (2) rtfm (2) rotfl (2) b4n (1) cu2 (1) irl (1) jff (1)
S2	dau (51) rl (9) linx (5) afk (4) rpg (2) rtfm (2) imho (2)
S3	rl (71) linx (20) imho (2) rtfm (1)

Tabelle 8: In S1 – S3 enthaltene Jargon-Ausdrücke und ihre jeweiligen Vorkommen

Technikbegeisterte Internet-Benutzer, viele Informatik-Studierende, Linux-Anwender, Benutzer von Chat-Systemen etc. setzen oftmals ein umfangreiches Inventar von Abkürzungen, Akronymen und speziellen Slangausdrücken ein. Um die drei Stichproben auf das Vorhandensein derartiger Ausdrücke und Abkürzungen zu untersuchen (einige, etwa „*gg*“ oder „*bg*“, befinden sich bereits in der Auflistung der Emphasen und werden an dieser Stelle somit nicht gesondert berücksichtigt, vgl. Tabelle 6), wurde eine etwa 1 000 Einträge umfassende Liste von Slangausdrücken[28] in das Analyseprogramm integriert. Da viele dieser Begriffe auch andere Bedeutungen haben können, wurden die jeweiligen Treffer manuell verifiziert, wobei sich etwa 120 Vorkommen als fehlerhaft herausgestellt haben.[29] Tabelle 8 zeigt die Vorkommen in den Stichproben. S1 enthält 90, S2 75 und S3 94 Slangausdrücke.

gewesen ist.

[28] Diese stammt, wie die zur automatischen Erkennung eingesetzte Smiley-Liste, von James Marshall, `http://www.astro.umd.edu/~marshall/`.

[29] Die Abkürzung ASAP beispielsweise kann neben „as soon as possible“ auch „Application Service Access Point“ bedeuten (vgl. `http://www.informatik.uni-leipzig.de/ifi/abteilungen/cs/dipl_txt/da_p_straeuber.html`).

Die in S1 enthaltenen Slangausdrücke gehen zurück auf eine alternative Version von „Dümmste anzunehmende User – Die Sammlung“ (`http://wwwstud.uni-giessen.de/~s5574/dau.htm`, 7 Vorkommen, insbesondere „DAU“), die bereits bei der Betrachtung der Iterationen erwähnten übersetzten Geschichten des „Bastard Operator from Hell“ (5 Vorkommen), eine Liste von „Linx, Linx, Linx“ (5 Vorkommen[30]) und eine weitere Übersetzung der BOFH-Legenden (`http://www.kawo1.rwth-aachen.de/~mo/texte/bofh4.html`, 4 Vorkommen). An der Spitze der Slangausdrücke enthaltenden Dokumente in S2 steht erneut „Dümmste anzunehmende User – Die Sammlung“ (`http://www.uni-giessen.de/~gc1091/jokes/dau.htm`, 48 Vorkommen, nicht mit o. g. Version zu verwechseln), gefolgt von zwei Kapiteln der bereits im Abschnitt über Smileys erwähnten Arbeit „Kommunikation und Rollenverhalten in Virtuellen Realitäten anhand des MUDs Xyllomer“ (5 bzw. 2 Vorkommen) und einer einzelnen Geschichte über einen ungeschickten Computerbenutzer, „Der Super-DAU! ([D]ümmster [A]nzunehmender [U]ser)“ (`http://www.uni-giessen.de/~gc1007/dau.html`, 1 Vorkommen). Die mit Abstand häufigsten Slangausdrücke in einem Dokument befinden sich in der bereits angesprochenen Arbeit „Kommunikationsstrukturen und Persönlichkeitsaspekte bei MUD-Nutzern“ (72 Vorkommen, fast ausschließlich „RL“, real life, einmal „IMHO“, in my humble opinion). Alle weiteren relevanten Dokumente aus S3 enthalten nur jeweils einen Slangausdruck, etwa eine E-Mail aus der bereits erwähnten TEX/LATEX Mailing-Liste[31] und Notizen zu einem Kurs über maschinelle Sprachverarbeitung an der Universität Heidelberg: Hier wird die Auflistung von Verweisen als „Linx“ markiert.

Auch bei dem Aspekt des Internet-Jargons setzt sich der Trend der primär in den studentischen Homepages herrschenden konzeptionellen Mündlichkeit fort. In dieser Stichprobe existieren mehr als 80 Dokumente, die ein – im Vergleich zu S2 und S3 – umfangreiches Vokabular von Slangausdrücken, Abkürzungen und Akronymen enthalten, das jedoch in puncto Quantität nicht mit der sehr extensiven Menge von Ausdrücken konkurrieren kann, die für Chatsysteme berichtet wurden (siehe Abschnitt 2). Die in S2 und S3 existierenden insges. 169 Vorkommen gehen primär auf die Dokumente „DAU – Die Sammlung“ und die beiden Arbeiten über MUD-Kommunikation zurück (zusammen 127 Vorkommen), woraus sich schließen lässt, dass Slangausdrücke sowohl in Dokumenten, die sich in tiefen Verzeichnissen auf Servern befinden, als auch in persönlichen Homepages von Universitätsangehörigen nur sehr selten benutzt werden, wohingegen in S1 eine recht breite Streuung von Belegen existiert.

4.4.6 Verschiedene weitere Merkmale

Dieser Abschnitt betrachtet einige weitere Merkmale, die tw. Hinweise auf den jeweiligen Grad der konzeptionellen Mündlichkeit einer Stichprobe liefern können: Die Belege von Bigraphen (z. B. „ue“ statt „ü“), Auslassungspunkten („. . . “), iterierten Interpunktionszeichen („!!!“) und Assimilationen von Wörtern („gibt's“). Die Ergebnisse dieser Analysen zeigt Tabelle 9.

Runkehl et al. (1998, S. 36 f.) haben festgestellt, dass die Verwendung von Bigraphen scheinbar „sehr stark mit der E-Mail“ korreliert: Es tauchen deutlich mehr dieser „Übertragungsfehler“ (ebd.) in E-Mails als in mit der Textverarbeitung oder der Schreibmaschine geschriebenen Briefen auf. Die jeweiligen Vorkommen habe ich mithilfe einer Suche nach Token, die Bigraphen enthalten, sowie einer sukzessive erstellten Negativliste von 12 715 Wörtern durchgeführt, die Sequenzen von „ae“, „ue“ und „oe“ enthalten, die jedoch nicht als Bigraphen zu werten sind (z. B. „aktuell“ oder „Abenteuer“). Die mit Abstand meisten Wörter, die Bigraphen enthalten, kommen in S3 vor (0,36% aller Token; S1: 0,12%, S2: 0,09%). Die Vorkommen von Bigraphen in S3 sind meiner Ansicht nach durch die Tatsache zu erklären, dass diese Stichprobe etwa 180 E-Mails[32] enthält, die aus HTML-Archiven von

[30] Siehe hierzu auch die Betrachtungen zum Soundalike-Slang in Haase et al. (1997, S. 73).

[31] „Wie Jens schon gesagt hat, liegt das Problem an den “Sichtbarkeitsregeln”, jedoch steckt IMHO der Fehler eigentlich an einer anderen Stelle: [. . .]“.

[32] Die HTML-Dokumente, in denen sich eingebettete E-Mails befinden, machen in S3 insgesamt 1,4% aller Token aus.

Merkmal	Studentische Homepages (S1)	Mitarbeiter-Homepages (S2)	Tiefe Dokumente (S3)
Bigraphen	10 190 Vork., 0,12% fuer (1017) koennen (172) zurueck (151) Universitaet (143) Jahrgaenge (96) natuerlich (93) wuerde (81) Fuer (76) waere (68) Zurueck (65)	4 413 Vork., 0,09% fuer (402) Muenchen (217) Universitaet (125) Buecher (60) Natuerliche (57) koennen (48) Joerg (38) Juergen (36) Jaeger (36) Einfuehrung (34)	11 305 Vork., 0,36% fuer (1250) koennen (119) Boerse (96) Gruenen (90) Muenchen (90) Fuer (85) Universitaet (82) zurueck (74) erklaerte (74) muesse (73)
Punkte	18 287 Vork., 0,21% ... (6073)	2 586 Vork., 0,05% ... (1775)	679 Vork., 0,02% ... (633)
Interpunktion	5 031 Vork., 0,06% !!! (1908) ‼ (1227) ??? (451) ?! (341) ⁇ (322) !? (196) !!!! (188) !!!!! (86) ???? (42) !!!!!! (35) ?!? (32) !!!!!!! (26) ????? (25) !?! (17) ?!! (15)	1 270 Vork., 0,03% !!! (446) ‼ (360) ⁇ (96) ??? (89) ?! (86) !!!! (46) !!!!! (23) !?! (21) ?!? (21) !? (16) ???? (15) !!? (6) ??????? (5) !!??? (4) !!!!!!! (4)	437 Vork., 0,01% !!! (138) ??? (83) ‼ (81) ⁇ (46) ?! (26) !!!!! (18) !!!! (12) !?! (9) ????? (6) ???? (3) !!!!!! (3) !? (2) ?????? (2) !!!!!!!! (1) ?!!! (1)
Assimilationen	13 048 Vork., 0,15% mit Großbst.: 5 326 mit Kleinbst.: 5 297 mit Apostroph: 2 425	4 203 Vork., 0,08% mit Großbst.: 2 285 mit Kleinbst.: 1 277 mit Apostroph.: 641	1 662 Vork., 0,05% mit Großbst.: 669 mit Kleinbst.: 418 mit Apostroph: 575
	I'm (106) Don't (90) Murphy's (88) It's (85) Arslandemir's (60) Bernd's (58) Chewy's (52) Can't (46) PC's (40) CD's (38)	Official's (122) Sa'dan (61) Wöll's (59) QBt's (54) B's (47) Attila's (44) Ferber'schen (41) A's (40) ATLt's (38) Drum'n (32)	Women's (23) CD's (16) Beck'sche (16) PC's (13) User's (12) Men's (11) AG's (9) Hill's (9) Bauer's (8) Labor's (8)
	gibt's (703) geht's (341) wird's (318) sieht's (242) gibtt's (193) don't (117) c't (117) war's (100) für's (99) ich's (87)	gibt's (93) geht's (55) c't (38) auf's (27) in's (24) gibtt's (21) d'un (21) don't (20) it's (19) d'imprimerie (16)	sieht's (21) gibt's (18) wird's (17) geht's (16) sprach's (12) don't (10) doesn't (10) c't (9) für's (9) d'enregistrements (8)

Tabelle 9: Bigraphen (die jeweils 10 häufigsten), Auslassungspunkte (das jeweils häufigste Vorkommen), iterierte Interpunktionszeichen (die 15 häufigsten) und Assimilationen (die jeweils 10 häufigsten Vorkommen mit initialem Groß- bzw. Kleinbuchstaben) in S1 – S3. Die nach den Vorkommen angegebenen Prozentzahlen sind deren prozentualer Anteil im Verhältnis zu den in einer Stichprobe enthaltenen Token.

Mailing-Listen stammen (genaueres hierzu in Abschnitt 4.5). Etwa bis zur Mitte der neunziger Jahre – etwa zu dieser Zeit etablierte sich der Standard MIME (Multipurpose Internet Mail Extensions), der eine Abhilfe für das im Folgenden skizzierte Problem lieferte – war es nicht problemlos möglich, Umlaute in E-Mails zu benutzen, da deren Übertragung an den Empfänger nicht sichergestellt werden konnte. Aus diesem Grund wurden zu dieser Zeit Zeichen, die Diakritika enthalten, mit Hilfe einer Umschrift emuliert, und diese Umschrift wird von vielen Benutzern des Mediums E-Mail auch heute noch aus Gründen der Gewöhnung eingesetzt, was deren häufiges Auftreten in S3 erklären könnte. Ein weiterer Ursprung könnten Dokumente wie die in der Analyse der Emphasen betrachtete FAQ-Liste der Newsgruppe `bln.markt` sein, die ursprünglich für das textbasierte Usenet verfasst wurde und möglicherweise aufgrund der sehr ähnlichen Produktionsbedingungen weitere Bigraphen enthält. Das prozentual etwas häufigere Auftreten von Bigraphen in S1 als in S2 lässt sich vermutlich ebenfalls durch Produktionsbedingungen erklären, da an einer Universität arbeitende Personen möglicherweise eher Wert darauf legen, dass ihre Dokumente fehlerfrei sind, als Studierende, deren Seiten oft nur ein Hobby und keine virtuelle Visitenkarte sind.

Auslassungspunkte („. . . “) werden von Storrer (2001a) als Zeichen für „das Innehalten in einem Turn“ im Rahmen der Chat-Kommunikation erklärt. Sicherlich haben Auslassungspunkte noch andere Funktionen, etwa das Signalisieren eines im Fluss befindlichen kognitiven Prozesses, dass der Autor einen Text nicht vollständig geplant, sondern hastig etwas aufgeschrieben hat. Die Vorkommen von Auslassungspunkten in S1 sind mit 18 287 (0,21% aller Token) im Vergleich zu S2 (0,05%) und S3 (0,02%) extrem[33] hoch, was darauf schließen lässt, dass viele Homepages von Studierenden ohne genaue Planung des Textes, jedoch mit dem Hintergedanken der direkten Kommunikation mit einer oder mehreren Personen, ähnlich rapide und unreflektiert produziert werden wie die E-Mail an einen Kommilitonen oder die einzelnen Zeilen eines flüchtigen, im Chat stattfindenden, Smalltalks. Weiterhin werden Auslassungspunkte auch häufig als Mittel der Sequenzbildung eingesetzt, wie etwa die jeweils mit „. . . “ beginnenden Bildunterschriften in der Galerie eines Skiurlaubs (`http://www.tu-harburg.de/~swps2910/home/champery.html`).

Im Usenet sind iterierte Interpunktionszeichen häufig Anlass, den Autor eines Artikels zu rügen (hierfür hat sich mittlerweile die Floskel „Deine !-Taste prellt.“ eingebürgert, vgl. auch Haase et al., 1997, S. 69), in den studentischen Homepages scheint dieses Mittel, eine wichtige Aussage zu betonen, jedoch gang und gäbe zu sein (S1: 0,06%; S2: 0,03%; S3: 0,01%). Diejenigen Dokumente mit den häufigsten Vorkommen sind Gästebücher, als HTML-Dokumente aufbereitete Chat-Protokolle und die bereits angesprochenen Sammlungen konvertierter Videotextseiten.

Runkehl et al. (1998) finden bei der Untersuchung verschiedener Korpora von E-Mails zwischen 2% und 4% Assimilationen (etwa „war's“), die zwar als „sprechsprachliche Mittel“ angesehen werden, jedoch nicht rekurrent auftauchen. Mit Hilfe eines regulären Ausdrucks filtert der in Perl implementierte Erkenner Vorkommen von Assimilationen, die ein Hochkomma enthalten und entweder mit einem Groß- oder Kleinbuchstaben oder einem Hochkomma beginnen.[34] Auch hier liegt S1 mit insgesamt 13 048 Vorkommen (0,15% aller Token) wieder vor S2 (0,08%) und S3 (0,05%). Interessant ist, dass „gibt's“ und „geht's“ in allen drei Stichproben zu den zwei (S1, S2) bzw. vier (S3) häufigsten Assimilationen gehören.

[33] Durch Vorkommen von Auslassungspunkten auf verschiedenen nach HTML konvertierte Videotextseiten wird dieser Statistik negativ beeinflußt.

[34] Nicht markierte Formen wie „wars“ können maschinell nur mit sehr viel Aufwand detektiert werden, weshalb derartige Vorkommen an dieser Stelle nicht in die Analyse einfließen können. Weiterhin ist die Filterung falscher Treffer wie etwa „don't“, „d'un“ oder „c't“ (eine Computerzeitschrift) maschinell schwierig zu realisieren. Da diese Treffer das Ergebnis nicht wesentlich beeinflussen, wurde auf eine manuelle Nachfilterung verzichtet.

4.5 In HTML-Dokumente eingebettete E-Mails und Newsartikel

In Abschnitt 4.4 wurden – etwa bei der Analyse der Bigraphen – HTML-Dokumente erwähnt, die E-Mails enthalten. Derartige Dokumente stammen fast ausschließlich aus im World Wide Web verfügbaren Archiven verschiedener Mailing-Listen. Zur Analyse der jeweiligen Vorkommen von in HTML-Dateien eingebetteten E-Mails und Newsartikeln wurde der Erkenner mit einer einfachen Detektionsheuristik ausgestattet, die ein Vorkommen meldet, falls die Zeichenketten „From:“, „Subject:“ und „Date:“ in einem Dokument enthalten sind.[35] Diese einfache Methode liefert bereits sehr zufriedenstellende Ergebnisse; bei einer (nicht systematisch durchgeführten) Untersuchung der als positiv markierten Dokumente wurden nur sehr wenige fehlerhafter Treffer entdeckt.

In S1 sind drei eingebettete E-Mails enthalten. Die erste E-Mail (`http://www.uni-ulm.de/~s_smasch/Moped_de/vergaser_bartz.html`) ist Teil einer Sammlung von Webseiten, die sich mit einem bestimmten Motorradtyp beschäftigen, und enthält Hinweise zur korrekten Einstellung des Vergasers. Hierbei ist unklar, ob dieser Beitrag ursprünglich aus einer Mailing-Liste oder einer Newsgruppe stammt. Die zweite E-Mail (`http://wwwstud.uni-giessen.de/~s6655/wetzlar.htm`) enthält Hinweise eines Internet-Providers zur korrekten Einstellung der Netzwerkoptionen unter MS Windows 95. Interessant ist hierbei, dass der Inhalt der E-Mail – mit Ausnahme der Absenderadressen – zwar in einer dicktengleichen Schrift dargestellt wird, einige Teile jedoch eingefärbt wurden. Auch hier ist der Ursprung der E-Mail unklar, es dürfte sich jedoch um eine direkte Antwort an den Autor der Homepage handeln. Bei der dritten E-Mail (`http://www.uni-ulm.de/~s_tfeger/selbst.html`) handelt es sich offenbar um eine Art Formschreiben, mit dem zur Selbstdarstellung neigende Betreiber von Homepages „abgemahnt“ werden. Der Empfänger dieser E-Mail und Autor der Homepage war hiervon offenbar so überrascht, dass er diese Mail in einer HTML-Version publiziert hat (das Dokument ist auf der Homepage durch den Link „Selbstdarsteller“ in der Sektion „Auf den Geist gehen mir“ erreichbar).

In S2 sind 10 eingebettete Newsartikel und zwei Sammlungen von E-Mails enthalten. Die beiden Sammlungen von E-Mails (und einigen offenbar redaktionellen Beiträgen) stammen aus einem dynamisch generierten „Pinboard“ (`http://www.uni-giessen.de/~gi63/Ebene3/pinboard.htm`), das von der Fachschaft Tiermedizin an der Universität Gießen betrieben wird und auf dem von dieser Fachschaft gesammelte Stellenangebote veröffentlicht werden. Einer der 10 Newsartikel („UKW Radiofrequenzen in Aachen“, `http://www-users.informatik.rwth-aachen.de/~guido/frequenzen.html`) erschien ursprünglich in der Newsgruppe `rwth.general`. Das Dokument wurde (offenbar manuell) tw. umformatiert, wobei einige typographische Möglichkeiten von HTML angewendet wurden. Die im Usenet und auch in E-Mails sehr häufig anzutreffende Signature (vgl. Schütte, 2000, S. 167 ff.) ist weiterhin in dem Dokument enthalten. Die verbleibenden 9 Newsartikel (etwa `http://userpage.chemie.fu-berlin.de/~biocheag/archiv/971201-vv-ein.html`) entstammen der Gruppe `bln.announce.fub.chemie`, in der Ankündigungen des Fachbereichs Chemie der FU Berlin publiziert werden. Diese Dokumente wurden nicht nachbearbeitet, sie erscheinen samt Header und Signature in einer Schreibmaschinenschrift, wie dies auch in einem herkömmlichen Newsreader der Fall wäre.[36]

[35] Nach „To:“ wird nicht gesucht, da die Adresse des Empfängers bei einem Mailing-Listen Archiv implizit bekannt ist und folglich nur selten angegeben wird. Handelt es sich um einen Newsartikel, wird die betroffene Gruppe in der Headerzeile „Newsgroups:“ notiert.

[36] Siehe Crowston und Williams (1999) für eine Analyse verschiedener Abstufungen derartiger Dokumente bei der Transformation von einem Medium (etwa Usenet) in ein Medium, das neue technische Möglichkeiten wie etwa Hyperlinks bietet. Ein hierzu treffendes Beispiel sind die an der TU Chemnitz online veröffentlichten Mitteilungen der Pressestelle. Ruft man eine solche, z. B. `http://www.tu-chemnitz.de/tu/presse/1998/05.14-08:33.html`, auf, erscheint eine „normale“, wenngleich recht schlichte Webseite. Am unteren Ende des Dokuments befindet sich jedoch eine klassische Signature, die – als einziger Bestandteil dieser Seite in einer Schreibmaschinenschrift gesetzt – die Straßenadresse, den Bearbeiter sowie verschiedene Kontaktadressen enthält. Die Adresse der Einstiegsseite der TU Chemnitz ist in der Signature mit einem Hyperlink versehen. Da man die jeweils neuesten Pressemitteilungen auch automatisch durch die Subskription einer Mailing-Liste erhalten kann, ist davon auszugehen, dass hier eine manuelle oder semi-automatische Übertragung der primär per E-Mail

S3 enthält schließlich 180 E-Mails und Newsartikel. Diese hohe Anzahl lässt sich dadurch erklären, dass Archive von Mailing-Listen nur sehr selten innerhalb von persönlichen Homepages abgelegt werden, sondern stattdessen in eher tiefen, offiziellen Bereichen eines Webservers zu finden sind, beispielsweise ein Beitrag der Ankündigungsliste des Gießener Hochschulrechenzentrums, `http://www.uni-giessen.de/hrz/hrznews/1996/msg00046.html` (hiervon insges. 15 E-Mails in S3). Archive von Mailing-Listen werden meist automatisch durch spezielle Programme erstellt, die die entstehenden HTML-Dokumente mit einer rudimentären Navigationsoberfläche versehen („vorherige", „nächste", „Index", „Thread"). Etwa 80 E-Mails stammen aus einem ebenfalls an der Universität Gießen verfügbaren HTML-Archiv der Mailing-Liste TEX-D-L.

4.6 Konstante Groß- oder Kleinschreibung

Nach Runkehl et al. (1998, S.36 f.) tritt in 7 bis 16% aller in verschiedenen Korpora von E-Mails enthaltenen Dateien eine konsequente Kleinschreibung auf, die vermutlich auf die oftmals schnelle und somit zeitsparende Produktionsweise zurückzuführen ist (vgl. u. a. Günther und Wyss, 1996, Quasthoff, 1997, Pansegrau, 1997). Da die automatische Erkennung von konstanter Groß- oder Kleinschreibung technisch einfach realisierbar ist, wurde ein entsprechendes Modul in das Analyseprogramm integriert. Die in einem Dokument enthaltene Anzahl von Wörtern wird hierbei jedoch nicht beachtet: Eine HTML-Datei, die lediglich ein Photo und darunter den mit einem Hyperlink versehenen Text „zurück" enthält, wird ebenfalls berücksichtigt.

Die studentischen Homepages (S1) enthalten insgesamt sechs Dokumente, die ausschließlich Groß- und 330 Dokumente, die ausschließlich Kleinbuchstaben enthalten. Die Homepages von Universitätsangehörigen (S2) enthalten zwei groß- und 23 kleingeschriebene Dateien. Die tiefen Dokumente (S3) umfassen 18 groß- und 53 kleingeschriebene Dokumente. Einige stichprobenartig betrachtete Webseiten zeigen, dass die meisten dieser Dateien tatsächlich nur Bildunterschriften enthalten, oder „Dummy"-Dokumente darstellen. Viele Dokumente scheinen Teile von Framesets zu sein, da ihr Zweck nicht auf den ersten Blick deutlich wird. Hierzu gehört z. B. eine in S3 enthaltene Seite, auf der lediglich „IRAN: WERBUNG, SPONSORING, AUSSCHREIBUNGEN" zu lesen ist. Im Hintergrund befindet sich eine schematische Darstellung der Umrisse des Iran, am unteren Ende befindet sich eine mit „Zurück zur Länderseite" beschriftete Grafik als Hyperlink. Vermutlich enthält dieses Dokument schlicht noch keinen Inhalt, wurde jedoch bereits vom Autor angelegt, um den eigentlichen Inhalt zu einem späteren Zeitpunkt einfacher einfügen zu können. Eine konstante Kleinschreibung wird auch in einigen HTML-Dokumenten (vor allem in denen der HDK-Berlin, beispielsweise `http://www.arch.hdk-berlin.de/projekte/venezia/venedig/10a_5.htm`, S3) als gestalterisches Stilmittel eingesetzt. Für vollständige Großschreibung scheint dies seltener der Fall zu sein.

4.7 Begrüßungen und Verabschiedungen

In der CMC-Literatur wurden insbesondere Begrüßungen und Verabschiedungen ausführlich untersucht (vgl. u. a. Lenke und Schmitz, 1995, Günther und Wyss, 1996, Haase et al., 1997, Pansegrau, 1997, Runkehl et al., 1998, Grzega, 1999). Im World Wide Web-Kontext hat sich de Saint-Georges (1998) im Rahmen einer Diskussion deiktischer Ausdrücke unter anderem mit Begrüßungen beschäftigt. Begrüßungsfloskeln wie „Welcome to my very own home-on-the web page" oder „Welcome to my fast-paced life" werden ihrer Ansicht nach eingesetzt, um das deiktische Zentrum – das virtuelle Zuhause – zu fixieren und den Leser der Seite, der sich bei der Lektüre einer Begrüßungsseite noch außerhalb dieses Zuhauses befindet, einzuladen: „Come on in and make yourself at home". Storrer (1999b) und Dürscheid (2000) sehen bei der Betrachtung des Lexems „Homepage" (Genaueres

verschickten Pressemitteilungen zur Zweitverwertung im WWW vorliegt.

zur Schreibweise dieses Wortes in Abschnitt 4.8) mit der kombinierten Benutzung eines räumlichen („home“) und eines bildlichen Ausdrucks („page“) einen „Metaphernmix“.

Zur maschinellen Untersuchung der drei Stichproben bzgl. der jeweiligen Vorkommen von Begrüßungen und Verabschiedungen wurden intuitiv zwei Listen mit üblichen Floskeln aufgestellt. Diese initialen Listen wurden einerseits durch in der CMC-Literatur genannte Begrüßungen und Verabschiedungen erweitert, andererseits durch die manuelle Analyse eines 1 000 Dokumente enthaltenen Samples privater Homepages. Insgesamt enthalten die Listen 39 verschiedene Begrüßungen und 42 Verabschiedungen.

S1 enthält insgesamt 3 287 Begrüßungen und 1 375 Verabschiedungen.[37] S2 enthält 1 005 Begrüßungs- und 634 Verabschiedungsfloskeln. S3 hingegen umfasst, wie an anderer Stelle bereits bemerkt wurde, eher Dokumente, die Inhalt transferieren sollen. Erwartungsgemäß finden sich hier lediglich 363 Begrüßungen und 228 Verabschiedungen.

Bei der Art und Weise der Begrüßungen ähneln sich die einzelnen Stichproben sehr. S1 enthält „Willkommen“ insgesamt 976 mal, gefolgt von „Hallo“ (824), „Welcome“ (450), „hi“ (278), „Herzlich Willkommen“ (270)[38], „Hey“ (235) und „Hello“ (54). Die sechs häufigsten Vorkommen in S2 lauten: „Willkommen“ (421), „Hallo“ (224), „Herzlich Willkommen“ (123), „Welcome“ (112), „Hi“ (31) und „Mahlzeit“ (31).[39] S3: „Willkommen“ (97), „Hallo“ (80), „Herzlich Willkommen“ (43), „Hi“ (41), „Hello“ (34) und „Welcome“ (22).

Die häufigsten Verabschiedungen fallen weniger homogen aus. S1: „Viel Spaß“ (332), „Grüße“ (193), „bis dahin“ (179), „Gruß“ (98), „CU“ (74) und „Ade“ (58). S2: „CU“ (220), „Viel Spaß“ (83), „bis dahin“ (73), „Gruss“ (46), „Grüße“ (34) und „Gruß“ (25). S3: „bis dahin“ (50), „Viel Spaß“ (36), „Grüße“ (20), „Gruß“ (19), „CU“ (17) und „Gruss“ (17).[40] Die Verabschiedungen scheinen insgesamt weniger konventionalisiert zu sein, jedoch machen die Listen deutlich, dass eher formelle Floskeln wie etwa „mit freundlichen Grüßen“ in *allen* Stichproben vollständig fehlen – vermutlich deshalb, weil deren Benutzung zu stark an die Brief- und E-Mail-Kommunikation gebunden ist und ihre Anwendung auf einer Webseite aufgrund des diesem Medium (d. h. der E-Mail-Kommunikation) primär inhärenten Dialogcharakters nicht angemessen erscheint.[41] Dennoch findet sich auf einigen studentischen Homepages ein Aufbau, der einer E-Mail bzw. einem herkömmlichen Brief sehr ähnlich ist, selbst ein Postscriptum ist gelegentlich enthalten (vgl. Abbildung 1).

4.8 „Homepage“ – „Home Page“ – „homepage“

Die verschiedensten, ihren Ursprung im Internet habenden, Begriffe finden sich mittlerweile in Wörterbüchern wieder. Bezüglich der Schreibung von „Homepage“ scheint, wenn man verschiedene Dokumente untersucht, die dieses Wort enthalten, gelegentlich Unklarheit zu herrschen. Das *Pons Großwörterbuch für Experten und Universität*[42] gibt das deutsche Wort als „Homepage“, die

[37] Es wird jeweils nur nach den entsprechenden Zeichenketten gesucht. Eine genauere Einschätzung, ob es sich bei einem Vorkommen von beispielsweise „Hallo“ wirklich um eine Begrüßung des Lesers handelt, könnte mit Hilfe einer Überprüfung der Position des Vorkommens innerhalb eines Dokuments (Anfang vs. Ende) realisiert werden.

[38] Bei der Erkennung einer Zeichenkette wie „Herzlich Willkommen“ wird lediglich hierfür ein Treffer gezählt; nur ein isoliertes Vorkommen von „Willkommen“ inkrementiert dessen Häufigkeit.

[39] Die häufigen Vorkommen von „Mahlzeit“ dürften aus dem Vorhandensein einiger Dutzend Dokumente aus einem ernährungswissenschaftlichen Institut resultieren.

[40] Nach Schütte (2000, S. 167) existiert auch häufig, u. a. in der Mailing-Liste INETBIB, das Ritual, die Grußformel mit einem „Wetterbericht“ zu flankieren: „Gruss aus Hamburg – schon wieder dunkel“. Im World Wide Web ist unklar, wann jemand ein Dokument lesen wird, daher erscheinen derartige, „Kopräsenz im gemeinsamen Wahrnehmungs- und Erfahrungsraum“ (ebd.) signalisierende Verabschiedungen und auch die von Quasthoff (1997) berichteten „tageszeitorientierten Grußformeln“ im World Wide Web intuitiv unangebracht und kommen in den drei Stichproben auch nicht vor.

[41] Feldweg et al. (1995, S. 147) berichten, dass „mfg“ als Abkürzung von „mit freundlichen Grüßen“ aufgrund „des verwendeten Briefstil[s]“ neben eher umgangssprachlichen Abschiedsformeln häufig in dem von ihnen untersuchten Korpus von mehr als 430 000 Newsartikeln auftauchen.

[42] Deutsch – Englisch, Englisch – Deutsch, Neubearbeitung (1999). Stuttgart, München etc.: Klett

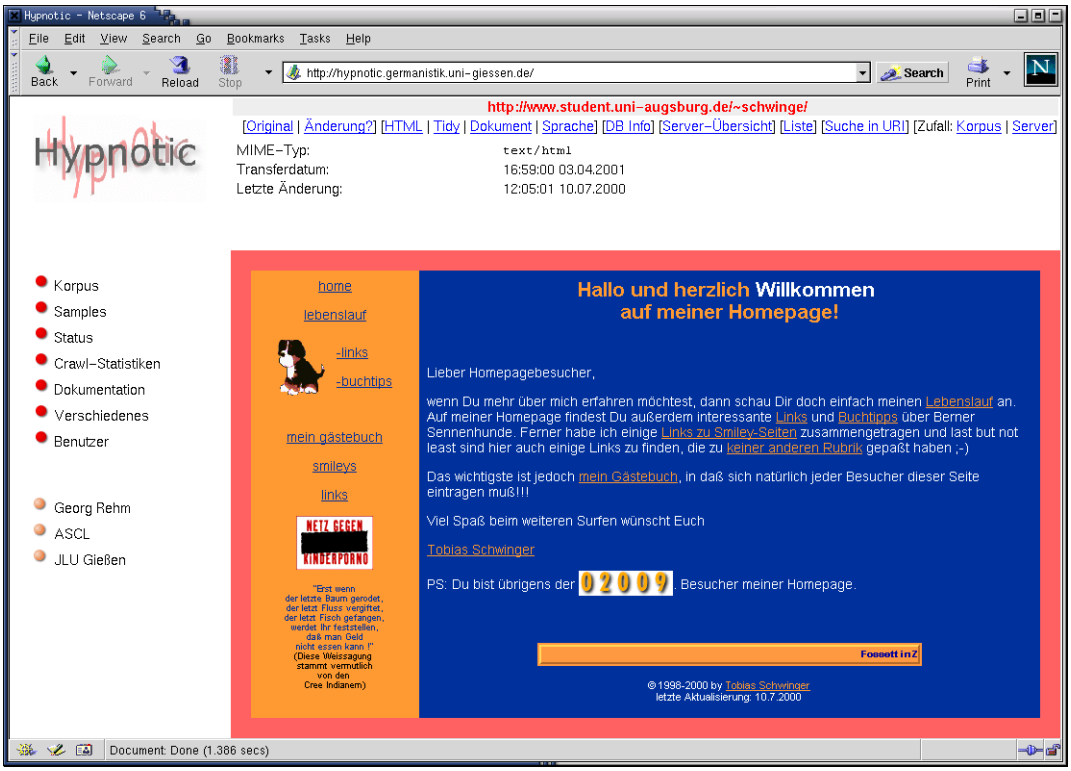

Abbildung 1: E-Mail-ähnlicher Aufbau einer studentischen Homepage

englische Übersetzung als „home page“ an. Letztere Schreibung ist konform mit derjenigen im *New Oxford Dictionary of Englisch* (Oxford University Press, 1998), die durch eine Konkordanzabfrage in der „Bank of English“ (`http://www.cobuild.collins.co.uk/boe_info.html`, enthält mehr als 30 Mio. Wörter) belegt wird: Dort werden die Schreibweisen „homepage“ oder „Homepage“ nicht gefunden, lediglich „home page“ (einmal „home-page“) ist in diesem Korpus enthalten.

Wie groß ist die Unklarheit tatsächlich? Um Varianten dieses neuen, jedoch bereits in Wörterbüchern etablierten Begriffs zu finden, wurde das Analyseprogramm um einen regulären Ausdruck erweitert, der alle Varianten des Begriffs „Homepage“ erkennt.[43] Insgesamt befinden sich 8 945 Vorkommen dieser Zeichenkette[44] in S1, 4 347 mal kommt der Begriff in S2 vor, S3 enthält schließlich lediglich 846 Vorkommen.

Die einzelnen Vorkommen selbst zeigen mit hoher Konsistenz, dass sich die Schreibweise „Homepage“ im deutschsprachigen Raum etabliert zu haben scheint. Die in den Stichproben enthaltenen,

[43] Analysen weiterer Schreibweisen sind denkbar: Wird vornehmlich „Website“, „website“ oder „Web-Site“ benutzt? Wie verhält es sich mit „E-Mail“, „Email“ oder „email“? Auch die häufig diskutierte Frage, ob es *die* oder *das E-Mail* heißt, könnte durch derartige Untersuchungen mit konkreten Daten beantwortet werden. Siehe hierzu auch Fußnote 4 in Pansegrau (1997, S. 86): „Trotz des inzwischen recht hohen Verbreitungsgrades der E-mail-Kommunikation existiert noch immer keine konventionalisierte Schreibweise: email, Email, e-mail, E-mail etc.“

[44] Alle Varianten eingeschlossen, jedoch nicht diejenigen, die in dem HTML-Element `<title>` auftauchen. Diese – beschränkt auf das genaue, jedoch Groß- und Kleinschreibung nicht beachtende Vorkommen von „Homepage“ – betragen für S1 2 284, für S2 nur 930 und für S3 lediglich 73 Dokumente. S1 enthält im Titel eines Dokuments folglich mit Abstand die meisten Vorkommen von „Homepage“. Die prozentualen Angaben über das Vorhandensein eines `<title>` Tags in den Stichproben lauten S1: 90,9%, S2: 95,7%, S3: 96,3%.

jeweils fünf häufigsten Belege sind: „Homepage“ (6698), „Homepages“ (489), „homepage“ (388), „Home Page“ (186), „HOMEPAGE“ (119). S2: „Homepage“ (3099), „Home Page“ (200), „ahs-homepage“ (162), „homepage“ (161), „Homepages“ (132). S3: „Homepage“ (582), „Home Page“ (38), „homepage“ (38), „Homepages“ (29), „[Home Page]“ (24). Bei den wenigen Vorkommen von „Home Page“ dürfte es sich um Entlehnungen aus englischsprachigen Dokumenten handeln, die aufgrund des nominalen Status der beiden Bestandteile in den hier betrachteten deutschsprachigen Seiten zwar mit initialen Großbuchstaben geschrieben werden, wobei der Einsatz der beiden Wörter als Kompositum jedoch nicht „gewagt“ wurde. Eine Untersuchung des Zeitstempels (des Datums der letzten Änderung) dieser Dokumente könnte Auskunft darüber geben, ob diese Dateien zuletzt vor einigen Jahren, als sich die Schreibweise „Homepage“ noch nicht durchgesetzt hatte, modifiziert wurden.

5 Schlussfolgerungen und Ausblick

Die in Abschnitt 4 dargestellten Analysen liefern durchaus positive Resultate, die die Gültigkeit der aufgestellten Hypothese andeuten – mit den „klassischen“ Kommunikationsdiensten des Netzes vertraute Studierende fügen die ihnen bekannten, Email-, IRC- und Usenet-spezifischen sprachlichen Phänomene in ihre privaten Homepages ein. S1, die Stichprobe studentischer Homepages, enthält beispielsweise deutlich mehr Smileys als die beiden anderen Samples (S1: 1 353 Vorkommen, 0,016% aller Token; S2: 298, 0,006%; S3: 93, 0,003%). Ebenso verhält es sich mit Iterationen (S1: 834, 0,010%; S2: 84, 0,002%; S3: 20, 0,001%) und Emphasen (S1: 182, 0,002%; S2: 38, 0,001%; S3: 24, 0,001%), wobei letztere insgesamt im World Wide Web weniger geläufig zu sein scheinen. Isolierte und nicht speziell durch Sonderzeichen hervorgehobene Verbstämme sind im World Wide Web sehr selten, wobei dieses Phänomen ausschließlich in S1 enthalten ist (S1: 25). Dies gilt auch für Slangausdrücke, sie kommen insgesamt nur selten vor, jedoch auch hier primär in S1 (S1: 90, S2: 75, S3: 94).[45] Der in S1 offenbar vorhandene saloppe, informelle Stil wird durch einen massiven Einsatz von Auslassungspunkten (S1: 18 287, 0,214%; S2: 2 586, 0,052%; S3: 679, 0,022%), reduplizierten Interpunktionszeichen (S1: 5 031, 0,059%; S2: 1 270, 0,026%; S3: 437, 0,014%) und Assimilationen (S1: 13 048, 0,153%; S2: 4 203, 0,085%; S3: 1 662, 0,053%) noch unterstrichen. Auch Begrüßungen (S1: 3 287, 0,039%; S2: 1 005, 0,020%; S3: 363, 0,012%) und Verabschiedungen (S1: 1 375, 0,016%; S2: 634, 0,013%; S3: 228, 0,007%) tauchen am häufigsten in S1 auf und kennzeichnen den sehr dialogbetonten Stil der enthaltenen Dokumente; dieses Ergebnis wird auch durch die Analyse der Wortfrequenzen gestützt. Weitere Untersuchungen sind jedoch notwendig, um diesen dialogbetonten Stil genauer zu untersuchen und mit der in Runkehl et al. (1998, S. 116) aufgestellten These – „Je stärker die Kommunikation dialogischer und synchroner erfolgt [E-Mail → Usenet → Chat, G. R.], desto häufiger lassen sich mündliche Aspekte des Sprachgebrauchs in der Internet-Kommunikation feststellen.“ – in Einklang zu bringen, schließlich wurde das World Wide Web nicht primär als *Kommunikations*medium geschaffen, sondern sollte vielmehr zum effektiven Verteilen von Information (Stichwort: Electronic Publishing) eingesetzt werden. Folglich befindet sich das Web in o. a. Sequenz der Internet-Kommunikation aus technischer Sicht deutlich jenseits der elektronischen Post, dennoch belegen die hier vorgestellten Analysen den massiven Einsatz sprechsprachlicher Merkmale in der Stichprobe studentischer Homepages. Der von den Autoren der betroffenen Dokumente offenbar präsupponierte Dialogcharakter – vgl. Abschnitt 2 – dieses Mediums könnte zum einen durch eine Unkenntnis und mangelnde Erfahrung mit der Produktion von Inhalten für das World Wide Web resultieren, zum anderen durch die Adaption und Transformation bekannter Kommunikationsdienste, beispielsweise E-Mail, so dass man hier meiner Ansicht

[45] Hierbei ist zu beachten, dass fast alle in S2 und S3 enthaltenen Belege aus den in Abschnitt 4 erwähnten wissenschaftlichen Arbeiten stammen, die sich mit der Internet-Kommunikation beschäftigen, es handelt sich also nicht um „authentische“ Belege.

nach – neben der Transformation von Merkmalen der konzeptionellen Mündlichkeit – auch von einer Textsorten-Transformation sprechen kann: Die in Abbildung 1 gezeigte studentische Homepage könnte man – bzgl. Inhalt und Form – durchaus als eine Art „generische E-Mail“ betrachten, die an alle potentiellen Leser dieser persönlichen Homepage gerichtet ist.

Runkehl et al. (1998, S. 116) sind der Auffassung: „Die sprachliche Variation zwischen den Diensten und innerhalb der Dienste ist – bei allen gemeinsamen rekurrenten Strukturen – deshalb besonders hervorzuheben, weil sich hier zeigt, daß Aussagen über ‘*die* Sprache in computervermittelter Kommunikation’, beziehungsweise über ‘*die* Sprache des Internet’ weit entfernt sind von der sprachlichen Realität, wie sich in ihrer Vielfältigkeit den Teilnehmern des Internet zeigt.“ Natürlich enthält nicht *jeder* im Internet veröffentlichte Text und nicht *jede* verschickte E-Mail *alle* in der Literatur beobachteten sprachlichen Phänomene, jedoch kann man durchaus von linguistischen Spezifika sprechen, die – mal sehr häufig, mal eher selten – in *allen* Kommunikationsdiensten des Internet von *sehr vielen* Benutzern bzw. Autoren eingesetzt werden. Eben dies war mit dem provokativen Schlagwort „Internetsprache“ in Haase et al. (1997) ursprünglich gemeint.

Neben der Darstellung der Ergebnisse besitzt dieser Beitrag einen weiteren Zweck: Die Motivation weiterer empirischer sprachwissenschaftlicher Arbeiten, die das World Wide Web als Kommunikationsmedium untersuchen. Die bisherigen Forschungen beschäftigen sich häufig auf einer sehr abstrakten Ebene (und mit proprietären und im Vergleich zum World Wide Web deutlich weniger häufig in der Benutzung befindlichen Systemen wie ToolBook oder HyperCard) mit Aspekten der Textkohärenz (etwa Storrer, 1999a, Fritz, 1999). Hier wäre es interessant, mittels Benutzerstudien und Korpusanalysen zu untersuchen, inwieweit die Kohärenz eines Webdokuments – etwa durch mangelnde Beschriftung von Hyperlinks – *tatsächlich* gefährdet ist oder ob überhaupt Kohärenzprobleme bei der alltäglichen Nutzung des Web bestehen. Dies hängt eng mit der Frage zusammen, wie HTML-Dokumente untereinander verknüpft werden: Oftmals wird das Konzept „Hypertext“ gleichsam als Heilsbringer[46] bezeichnet, das Aufbrechen der Linearität fördere Lernprozesse und rege zu neuen Assoziationen an. Meiner Ansicht nach existiert eine solche, in der klassischen Hypertext-Theorie verankerte, umfangreiche Vernetzung de facto nicht. Es gibt im World Wide Web nur sehr wenige Hypertexte, die über sehr viele einzelne Textknoten verfügen und die untereinander sehr extensiv miteinander vernetzt sind, da der Aufwand, solche Texte entsprechend aufzubereiten, sehr groß ist und m. E. der alltägliche Benutzer keinen Bedarf hierfür hat und die Kenntnis des *eigentlichen* Hypertext-Konzepts nur bei sehr wenigen Autoren vorhanden ist.[47] Üblicher hingegen sind in sich abgekapselte Texte, die lediglich mittels „vor“, „zurück“ und „nach oben“ – monosequenziert im Sinne von Storrer (2000b) – zu rezipieren sind. Natürlich werden nur wenige Seiten „isoliert ins Web gestellt, enthalten also keine Verknüpfungen, keine Links auf andere Seiten“ (Dürscheid, 2000, S. 62), jedoch darf man die Tatsache, dass eine Webseite einen oder mehrere Verweise auf eine andere Seite enthält, nicht gleichsetzen mit der Implementierung und Realisierung des Hypertext-Konzepts, wie es – angefangen bei Bush (1945) – ab der Mitte der sechziger Jahre vornehmlich von Ted Nelson (1987) entwickelt wurde (vgl. auch die umfangreiche Bibliographie in Kuhlen, 1991). Im Übrigen existieren hierzu beim derzeitigen Stand der Technik – trotz XML (Bray et al., 2000) und flankierender Standards – noch längst nicht alle technischen Bedingungen, vgl. Bieber et al. (1997) oder Furuta und Marshall (1996): „The hypermedia research community often views the Web with a combination of awe and frustration. Awe because of its meteoric ascendance [...]. Frustration because of the sensation that hard lessons learned with hypermedia are not being reflected.“

Die in diesem Beitrag präsentierte Studie ist Teil eines Forschungsprojekts, das sich mit der Untersuchung von Hypertextsorten beschäftigt. Hierbei sollen die Fragen geklärt werden, welche Hypertextsorten im World Wide Web existieren, welche Merkmale sie aufweisen und mit welchen Metho-

[46] Siehe beispielsweise Dillon (1996) für kritische Anmerkungen.

[47] Wie die Analysen in Abschnitt 4 gezeigt haben, spielen auch die Produktionsbedingungen eines Textes – wurde er beispielsweise originär für eine Newsgruppe geschrieben und dann lediglich im WWW einer Zweit- oder Alternativverwertung unterzogen – bei dieser Diskussion eine sehr große Rolle (hierzu auch Storrer, 2000a).

den sie maschinell klassifizierbar sind. Die Ergebnisse der Studie sind bezüglich dieses Vorhabens sehr vielversprechend, so könnte man beispielsweise den „Grad der konzeptionellen Mündlichkeit“ einer Webseite anhand der hier untersuchten sprachlichen Spezifika berechnen und in eine Klassifikationskomponente als eines von vielen weiteren Merkmalen einbeziehen. Hierzu gehören beispielsweise die jeweilige Art der Vernetzung einer Gruppe von Dokumenten, Dokumentlängen, Anzahl und Formate der eingebetteten Graphiken, der Einsatz von Formularen, Java Applets und JavaScript (vgl. Rehm, 2002). Das langfristige Ziel ist, die theoretischen und praktischen Grundlagen einer robusten Klassifikation von HTML-Dokumenten in ihre jeweiligen Hypertextsorten zu etablieren, um somit den Benutzern ein weiteres Werkzeug zur Verfügung zu stellen, den Informationsdschungel World Wide Web effizienter nutzen zu können.

Literatur

Amitay, Einat (2000): “Anchors in Context: A Corpus Analysis of Web Pages Authoring Conventions”. In: *Words on the Web*, herausgegeben von Pemberton, Lynn und Shurville, Simon, Bristol: Intellect Books, S. 25–35. Online verfügbar: `http://www.mri.mq.edu.au/~einat/`. Die Printfassung dieses Textes wurde gekürzt; Zitate beziehen sich auf die online erschienene Version.

Angell, David und Heslop, Brent (1994): *The Elements of E-mail Style*. Reading, Menlo Park, New York etc.: Addison-Wesley.

Beisswenger, Michael (Herausgeber) (2001): *Chat-Kommunikation. Sprache, Interaktion, Sozialität & Identität in synchroner computervermittelter Kommunikation. Perspektiven auf ein interdisziplinäres Forschungsfeld*. Stuttgart: ibidem. Im Druck.

Berners-Lee, Tim; Cailliau, Robert; Groff, Jean-François und Pollermann, Bernd (1992): “World-Wide Web: The Information Universe”. *Electronic Networking: Research, Applications and Policy* 1 (2).

Bieber, Michael; Vitali, Fabio; Ashman, Helen; Balasubramanian, V. und Oinas-Kukkonen, Harri (1997): “Fourth generation hypermedia: Some missing links for the world wide web”. *International Journal of Human-Computer Studies* 47: S. 31–65.

Bray, Tim; Paoli, Jean; Sperberg-McQueen, C. M. und Maler, Eve (2000): “Extensible Markup Language (XML) 1.0 (Second Edition)”. Technische Spezifikation, World Wide Web Consortium. Online verfügbar: `http://www.w3.org/TR/2000/REC-xml-20001006`.

Bucher, Hans Jürgen (1996): “Textdesign – Zaubermittel der Verständlichkeit? Die Tageszeitung auf dem Weg zum interaktiven Medium”. In: *Textstrukturen im Medienwandel*, herausgegeben von Hess-Lüttich, Ernest W. B.; Holly, Werner und Püschel, Ulrich, Frankfurt am Main, Berlin etc.: Lang, Band 29 von *Forum Angewandte Linguistik*, S. 31–59.

Bush, Vannevar (1945): “As we may think”. *Atlantic Monthly* 176 (1): S. 101–108.

Crocker, David H. (1982): “Standard for the Format of ARPA Internet Text Messages”. Network Working Group – Request for Comments (RFC) 822. Online verfügbar: `http://www.rfc-editor.org`.

Crowston, Kevin und Williams, Marie (1999): “The Effects of Linking on Genres of Web Documents”. In: *Proceedings of the 32nd Hawaii International Conference on Systems Sciences (HICSS-32)*. IEEE.

de Saint-Georges, Ingrid (1998): “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.

Diekmannshenke, Hajo (2000): “Die Spur des Internetflaneurs – Elektronische Gästebücher als neue Kommunikationsform”. In: *Soziales im Netz – Sprache, Beziehungen und Kommunikationskulturen im Internet*, herausgegeben von Thimm, Caja, Opladen, Wiesbaden: Westdeutscher Verlag, S. 131–155.

Dillon, Andrew (1996): “Myths, Misconceptions, and an Alternative Perspective on Information Usage and the Electronic Medium”. In: *Hypertext and Cognition*, herausgegeben von Rouet, Jean-Francois; Levonen, Jarmo J.; Dillon, Andrew und Spiro, Rand J., Mahwah: Erlbaum, S. 25–42.

Dürscheid, Christa (1999): “Zwischen Mündlichkeit und Schriftlichkeit”. *Papiere zur Linguistik* 60 (1): S. 17–30.

Dürscheid, Christa (2000): “Sprachliche Merkmale von Webseiten”. *Deutsche Sprache* 28 (1): S. 60–73.

Feldweg, Helmut; Kibiger, Ralf und Thielen, Christine (1995): “Zum Sprachgebrauch in deutschen Newsgruppen”. *Osnabrücker Beiträge zur Sprachtheorie* (50): S. 143–154.

Fielding, R.; Gettys, J.; Mogul, J. C.; Frystyk, H.; Masinter, L.; Leach, P. und Berners-Lee, T. (1999): “Hypertext Transfer Protocol – HTTP/1.1”. Network Working Group – Request for Comments (RFC) 2616. Online verfügbar: `http://www.rfc-editor.org`.

Friedl, Jeffrey E. F. (1997): *Mastering Regular Expressions*. Cambridge, Köln, Paris etc.: O’Reilly & Associates.

Fritz, Gerd (1999): “Coherence in Hypertext”. In: *Coherence in Spoken and Written Discourse*, herausgegeben von Bublitz, Wolfram; Lenk, Uta und Ventola, Eija, Amsterdam, Philadelphia: John Benjamins, Nummer 63 in Pragmatics And Beyond New Series, S. 221–232.

Furuta, Richard und Marshall, Catherine C. (1996): “Genre as Reflection of Technology in the World-Wide Web”. Technischer Bericht, Hypermedia Research Lab, Department of Computer Science, Texas A&M University.

Gruber, Helmut (1997): “Themenentwicklung in wissenschaftlichen E-mail-Diskussionslisten. Ein Vergleich zwischen einer moderierten und einer nichtmoderierten Liste”. In: *Sprachwandel durch Computer*, herausgegeben von Weingarten, Rüdiger, Opladen: Westdeutscher Verlag, S. 105–128.

Grzega, Joachim (1999): “Some Observations on E-Mail Style vs. Traditional Style”. *Papiere zur Linguistik* 60 (1): S. 3–16.

Günther, Ulla und Wyss, Eva Lia (1996): “E-mail-Briefe – eine neue Textsorte zwischen Mündlichkeit und Schriftlichkeit”. In: *Textstrukturen im Medienwandel*, herausgegeben von Hess-Lüttich, Ernest W. B.; Holly, Werner und Püschel, Ulrich, Frankfurt am Main, Berlin etc.: Lang, Band 29 von *Forum Angewandte Linguistik*, S. 61–86.

Haase, Martin; Huber, Michael; Krumeich, Alexander und Rehm, Georg (1997): “Internetkommunikation und Sprachwandel”. In: *Sprachwandel durch Computer*, herausgegeben von Weingarten, Rüdiger, Opladen: Westdeutscher Verlag, S. 51–85.

Koch, Peter und Oesterreicher, Wulf (1994): “Schriftlichkeit und Sprache”. In: *Schrift und Schriftlichkeit*, herausgegeben von Günther, H. und Ludwig, O., Berlin, New York: de Gruyter, Band 1 von *Handbücher für Sprach- und Kommunikationswissenschaft*, S. 587–604.

Kuhlen, Rainer (1991): *Hypertext – Ein nicht-lineares Medium zwischen Buch und Wissensbank*. Berlin, Heidelberg, New York etc.: Springer.

Lenke, Nils und Schmitz, Peter (1995): “Geschwätz im ‚Globalen Dorf’ – Kommunikation im Internet”. *Osnabrücker Beiträge zur Sprachtheorie* (50): S. 117–141.

Nelson, Theodor Holm (1987): “Literary Machines”. Eigenverlag. Edition 87.1.

Pansegrau, Petra (1997): “Dialogizität und Degrammatikalisierung in E-mails”. In: *Sprachwandel durch Computer*, herausgegeben von Weingarten, Rüdiger, Opladen: Westdeutscher Verlag, S. 86–104.

Quasthoff, Uta M. (1997): “Kommunikative Normen im Entstehen: Beobachtungen zu Kontextualisierungsprozessen in elektronischer Kommunikation”. In: *Sprachwandel durch Computer*, herausgegeben von Weingarten, Rüdiger, Opladen: Westdeutscher Verlag, S. 23–50.

Raggett, Dave; Hors, Arnaud Le und Jacbos, Ian (1999): “HTML 4.01 Specification”. Technische Spezifikation, World Wide Web Consortium. Online verfügbar: `http://www.w3.org/TR/html401/`.

Rehm, Georg (2001): “`korpus.html` – Zur Sammlung, Datenbank-basierten Erfassung, Annotation und Auswertung von HTML-Dokumenten”. In: *Proceedings of the GLDV Spring Meeting 2001*, herausgegeben von Lobin, Henning. Gesellschaft für linguistische Datenverarbeitung (Society for Computational Linguistics and Language Technology), Giessen, Germany, S. 93–103. Online verfügbar: `http://www.uni-giessen.de/fb09/ascl/gldv2001/`.

Rehm, Georg (2002): “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*. Big Island, Hawaii: IEEE. Im Druck.

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

Sassen, Claudia (2000): “Phatische Variabilität bei der Initiierung von Internet-Relay-Chat-Dialogen”. In: *Soziales im Netz – Sprache, Beziehungen und Kommunikationskulturen im Internet*, herausgegeben von Thimm, Caja, Opladen, Wiesbaden: Westdeutscher Verlag, S. 89–108.

Schlobinski, Peter (2000): “Chatten im Cyberspace”. In: *Die deutsche Sprache zur Jahrtausendwende*, herausgegeben von Eichhoff-Cyrus, Karin M. und Hoberg, Rudolf, Mannheim, Leipzig, Wien, Zürich: Dudenverlag, S. 63–79.

Schmidt, Gurly (2000): “Chat-Kommunikation im Internet – eine kommunikative Gattung”. In: *Soziales im Netz – Sprache, Beziehungen und Kommunikationskulturen im Internet*, herausgegeben von Thimm, Caja, Opladen, Wiesbaden: Westdeutscher Verlag, S. 109–130.

Schütte, Wilfried (2000): “Sprache und Kommunikationsformen in Newsgroups und Mailinglisten”. In: *Sprache und neue Medien. Jahrbuch des Instituts für deutsche Sprache 1999*, herausgegeben von Kallmeyer, Werner, Berlin, New York: de Gruyter, S. 142–178.

Storrer, Angelika (1999a): “Kohärenz in Text und Hypertext”. In: *Text im digitalen Medium – Linguistische Aspekte von Textdesign, Texttechnologie und Hypertext Engineering*, herausgegeben von Lobin, Henning, Wiesbaden: Westdeutscher Verlag, S. 33–65.

Storrer, Angelika (1999b): “Was ist eigentlich eine Homepage? Neue Formen der Wissensorganisation im World Wide Web”. *Sprachreport* (1): S. 2–8. Online verfügbar: `http://www.ids-mannheim.de/grammis/storrer.html`.

Storrer, Angelika (2000a): “Schriftverkehr auf der Datenautobahn: Besonderheiten der schriftlichen Kommunikation im Internet”. In: *Neue Medien im Alltag*, herausgegeben von Voß, G. Günther; Holly, Werner und Boehnke, Klaus, Opladen: Leske + Budrich, S. 151–175.

Storrer, Angelika (2000b): “Was ist „hyper“ am Hypertext?” In: *Sprache und neue Medien. Jahrbuch des Instituts für deutsche Sprache 1999*, herausgegeben von Kallmeyer, Werner, Berlin, New York: de Gruyter, S. 222–249.

Storrer, Angelika (2001a): “Getippte Gespräche oder dialogische Texte? Zur kommunikationstheoretischen Einordnung der Chat-Kommunikation”. In: *Sprache im Alltag. Beiträge zu neuen Perspektiven der Linguistik*, Berlin etc.: de Gruyter, S. 439–466. Herbert Ernst Wiegand zum 65. Geburtstag gewidmet.

Storrer, Angelika (2001b): “Neue Medien – neue Stilfragen: Das World Wide Web unter stilistischer Perspektive”. In: *Perspektiven auf Stil*, herausgegeben von Jakobs, Eva-Maria und Rothkegel, Annely, Tübingen: Niemeyer. Im Druck.

van Berkel, Arrie und de Jong, Mariët (1999): “Coherence Phenomena in Hypertextual Environments”. In: *Textproduktion: HyperText, Text, KonText*, herausgegeben von Jakobs, Eva-Maria; Knorr, Dagmar und Pogner, Karl-Heinz, Frankfurt am Main, Berlin, Bern etc.: Lang, Band 5 von *Textproduktion und Medium*, S. 29–40.

Wall, Larry; Christiansen, Tom und Orwant, Jon (2000): *Programming Perl*. Cambridge, Köln, Paris etc.: O’Reilly & Associates, 3. Auflage.

Ziegler, Arne (2001): “Zur @kronymischen Verwendung der Phraseologismen in Textsorten der Internet-Kommunikation am Beispiel der E-Mail”. In: *Wer A sägt, muss auch B sägen. Phraseologie und Parömiologie*, herausgegeben von Hartmann, Dietrich und Wirrer, Jan, Hohengehren: Schneider. Im Druck.