Thomas Bartz, Michael Beißwenger, Angelika Storrer

Optimierung des Stuttgart-Tübingen-Tagset für die linguistische Annotation von Korpora zur internetbasierten Kommunikation: Phänomene, Herausforderungen, Erweiterungsvorschläge

1 Einleitung

Die Erschließung, Dokumentation und Aufbereitung von Sprachdaten aus Genres internetbasierter Kommunikation für die Zwecke der korpusgestützten empirischen Sprachanalyse stellt gegenwärtig eine große Herausforderung für den Bereich der Korpuslinguistik und Sprachbeschreibung dar (Beißwenger & Storrer 2008): Die Besonderheiten der schriftlichen Sprachverwendung in E-Mails, Chats, Online-Foren, Twitter, Wikipedia-Diskussionsseiten, Weblog-Kommentaren, sozialen Netzwerken, Instant-Messaging-Anwendungen oder Online-Computerspielen (MMORPGs) lassen sich, wie die neuere linguistische Forschung in diesem Bereich gezeigt hat, weder mit Kategorien und Modellen für die Analyse geschriebener Texte noch mit Kategorien und Modellen für die Analyse mündlicher Gespräche zufriedenstellend beschreiben. Auch computerlinguistische Verfahren für die automatische Analyse und Annotation geschriebener Sprache lassen sich zum gegenwärtigen Stand nur sehr bedingt für die Aufbereitung von Sprachdaten aus Genres internetbasierter Kommunikation heranziehen. Darüber hinaus existieren bis dato keinerlei Standards für die Strukturannotation und Repräsentation solcher Daten in Korpora. Korpora internetbasierter Kommunikation stellen neben den Text- und Gesprächskorpora einen „Korpustyp dritter Art“ (Storrer i.Dr.) dar, für deren Aufbau geeignete Standards, Verfahren und Gütekriterien erst noch entwickelt werden müssen.

Weshalb die Sprachverwendung in der internetbasierten Kommunikation neuer, dem Untersuchungsgegenstand angemessener Beschreibungs- und Analyseansätze bedarf, wurde in der linguistischen Forschung zum Thema bereits ausführlich theoretisch begründet und empirisch gezeigt (vgl. z.B. Herring 1996, 1999, 2010/11; Beißwenger 2000, 2007; Crystal 2001, Storrer 2001, Storrer i.Dr.; Bittner 2003; Schönfeldt & Golato 2003). Die entsprechenden Befunde sind aber bislang kaum in korpus- und computerlinguistische Ansätze zur Behandlung von Sprachdaten aus diesem Kommunikationsbereich eingeflossen. Erste Schritte, die linguistische Theoriebildung zu diesem Bereich mit korpus- und computerlinguistischen Beschreibungsansätzen zusammenbringen, werden gegenwärtig im europäischen Netzwerk „Building and Annotating CMC Corpora“[1] sowie in der Special Interest Group „Computer-Mediated Communication“ im Rahmen der *Text Encoding Initiative* (*TEI*)[2] unternommen; ein erster Entwurf zu einem Annotationsschema für Genres internetbasierter Kommunikation, das linguistische Ergebnisse zum Gegenstand mit einem existierenden Repräsentationsstandard im Bereich der E-Humanities zusammenführt, ist in Beißwenger et al. (2012) beschrieben.

Werkzeuge und Verfahren für die automatische Wortartenannotation (im Folgenden auch: *Part-of-speech-Tagging*, *POS-Tagging*) stellen neben Repräsentationsstandards ein weiteres

grundlegendes Desiderat beim Aufbau linguistisch aufbereiteter Korpora zur Sprachverwendung in der internetbasierten Kommunikation dar (Beißwenger & Storrer 2008: 302f., King 2009: 312f.): Existierende Verfahren, die in der Regel auf Korpora mit redigierten Texten (Zeitungstexten u.Ä.) trainiert wurden, lassen sich zum gegenwärtigen Stand der Kunst nicht ohne deutliche Einbußen bei der Genauigkeit der Klassifikation auf schriftliche Sprachdaten aus E-Mails, Chats, Online-Foren, Twitter, Wikipedia-Diskussionsseiten, Weblog-Kommentaren, sozialen Netzwerken oder Instant-Messaging-Anwendungen anwenden.

Giesbrecht & Evert (2009) zeigen in einem systematischen Vergleich der Performanz verschiedener Wortartentagger, dass sich die Zuverlässigkeit der Wortartenzuordnung drastisch verschlechtert, wenn anstelle von Texten mit redigierter Schriftlichkeit Webkorpora verarbeitet werden. Korpusausschnitte mit geringerer Konformität zum geschriebenen Standard (z.B. Postings aus Online-Foren, die Giesbrecht & Evert entsprechend als „hard genres" bezeichnen) verursachen dabei größere Probleme als Dokumente mit höherer Standardkonformität („easy genres"). Bick (2010) kommt zu ähnlichen Befunden; als Beispiele für Phänomene, die in der E-Mail- und in noch stärkerem Maße in der Chat-Kommunikation Verarbeitungsprobleme verursachen, benennt er u.a. kontraktierte Formen (*dont*, *gotta*), Fälle der graphischen Nachbildung von Phänomenen der gesprochenen Sprache („phonetic writings": *Ravvvvvvvvvvvvveeee*), netztypische Akronyme sowie „subject-less sentences" wie *dances [around] wild and naked*. Weitere Herausforderungen für die automatische Analyse ergeben sich durch Schnellschreibphänomene (Tippfehler, Wortauslassungen, Verzicht auf normkonforme Großschreibung) sowie durch Einheiten, die keine direkten Pendants in Genres redigierter Schriftlichkeit haben; zu nennen wären hier u.a. Emoticons, Adressierungen (*@bienchen*, *@alle*) sowie Aktionswörter wie *grins*, *lach*, *lol*, *stirnrunzel* oder *diestirninfaltenleg*.

Mit existierenden, für linguistische Anwender direkt nutzbaren Verarbeitungswerkzeugen können Phänomene dieser Art bislang nur unzureichend analysiert und disambiguiert werden. Entsprechend sind korpusgestützte Analysen zu den sprachlichen Besonderheiten und zur sprachlichen Variation in Genres internetbasierter Kommunikation bislang noch mit großem Aufwand verbunden: Entweder kann sich die Analyse nur auf Rohdaten oder auf einige wenige in die Daten eingebrachte Annotationen stützen oder die verwendeten Korpora müssen für Analysezwecke vorab zeit- und kostenintensiv handannotiert werden.

Inzwischen gibt es zumindest für das Englische und das Niederländische allerdings erste Arbeiten, in denen Verfahren zur Wortartenannotation auf einzelne Genres internetbasierter Kommunikation angepasst wurden. So wurden in Ritter et al. (2011), Gimpel et al. (2011) und Owoputi et al. (2012) Wortarten-Tagsets und darauf bezogene Verarbeitungswerkzeuge für die Annotation englischer Twitter-Daten optimiert; die zugrunde gelegten Tagsets enthalten Kategorien für Hashtags, Adressierungen („@-mentions"), Emoticons, URLs und Retweet-Marker (*RT*), bei Gimpel et al. (2011) auch für konktraktierte Formen wie *someone's*, *I'm*, *let's*, *book'll*, *Mark'll*. Avontuur et al. (2012) nutzen die in Gimpel et al. (2011) beschriebenen Twitter-spezifischen Tagset-Erweiterungen in Kombination mit einem existierenden Tagset für das Niederländische für die Annotation niederländischer Tweets. Die Ergebnisse der auf der Grundlage dieser Tagsets entwickelten automatischen Verfahren

sind ermutigend; die Entwicklung vergleichbarer Verfahren für das Deutsche und unter Einbeziehung auch anderer Genres als nur Twitter sowie weiterer sprachlicher Phänomene ist aber noch zu leisten. Speziell mit Blick auf die korpusgestützte linguistische Analyse sprachlichen Wandels durch internetbasierte Kommunikation ist es zudem wünschenswert, netztypische sprachliche Innovationen nicht einfach als neue, „online-spezifische" (Gimpel et al. 2011: 2) Kategorien in die verwendeten Tagsets einzufügen, sondern sie in einen wortartentheoretischen Beschreibungsrahmen zu integrieren und – wo begründbar – als online-spezifische Erweiterungen vorhandener Kategorien darzustellen. In Beißwenger et al. (2012: 3.5.1) wird die Möglichkeit einer solchen Integration für Emoticons, Aktionswörter und Adressierungen aufgezeigt.

Welche Potenziale sich der Variations- und Medienlinguistik durch die Analyse großer annotierter Korpora zur internetbasierten Kommunikation bieten, zeigen exemplarisch die Untersuchungen von Bick (2010) zu Mündlichkeitsphänomenen und grammatischen Merkmalen in einem E-Mail- und Chat-Korpus mit insgesamt 117 Millionen Tokens. An welche Kapazitätsgrenzen man stößt, wenn man die Analyse eines großen Korpus zur internetbasierten Schriftlichkeit ohne die Möglichkeit zur Suche über linguistischen Kategorien (Annotationen) durchführen muss, belegt die quantitative Untersuchung von Storrer (2013) zu Frequenzunterschieden bei der Verwendung von Emoticons auf den Artikel- und Diskussionsseiten der deutschen Wikipedia: Obwohl anhand formaler Merkmale recht gut per Volltextsuche identifizierbar, gibt es in der Wikipedia zu Emoticons homonyme Zeichenkombinationen, die selbst keine Emoticons sind. Je fünfstellige Trefferzahlen für die Emoticon-Formen *;-), :-)* und *:)* lassen sich aber intellektuell nicht mehr mit vertretbarem Aufwand disambiguieren; entsprechend ist in solchen Fällen eine Differenzierung nach falsch positiven Treffern (Pseudotreffern) und echten Belegen ohne die Möglichkeit zur Suche über Kategorien nicht mehr möglich. Emoticons stellen hierbei noch einen vergleichsweise „einfachen" Fall dar; für die empirische Analyse komplexerer sprachlicher Phänomene dürften, wenn lediglich die Möglichkeit einer Volltextsuche zur Verfügung steht, die methodischen Probleme noch ungleich größer sein.

Linguistisch annotierte Korpora mit Sprachdaten aus Genres internetbasierter Kommunikation stellen dabei nicht nur ein Desiderat desjenigen Bereichs variations- und medienlinguistischer Forschung dar, der sich speziell mit den Besonderheiten der Sprachverwendung und Kommunikation im Internet beschäftigt; auch empirische Untersuchungen zur Beschreibung und Analyse von aktuellen Tendenzen in der Entwicklung der deutschen Gegenwartssprache kommen im Zeitalter digitaler Kommunikationstechnologien nicht umhin, die internetbasierte Kommunikation als einen wichtigen und aktiven Bereich sprachlicher Variation und sprachlichen Wandels einzubeziehen und gestützt auf authentische Daten zu untersuchen. Nicht zuletzt werden Verfahren für die automatische linguistische Analyse von Sprachdaten aus Genres internetbasierter Kommunikation auch im Bereich der Sprachtechnologie benötigt, wenn es darum geht, große, aus dem World Wide Web erhobene „Webkorpora" mit automatischen Methoden für die Nutzung in sprachtechnologischen Anwendungen aufzubereiten und zu analysieren.

Um die Wortartenannotation für Sprachdaten aus Genres internetbasierter Kommunikation zu verbessern, benötigt man Kategorien, Ressourcen und Verfahren auf unterschiedlichen Ebenen:

(i) Kategorien und linguistische Beschreibungen zu charakteristischen Phänomenen internetbasierter Schriftlichkeit, die bei der Verarbeitung mit auf redigierten Texten trainierten Verarbeitungswerkzeugen typischerweise Probleme verursachen;

(ii) eine Typologie der Probleme, die sich bei der Behandlung dieser Phänomene auf unterschiedlichen Ebenen des Verarbeitungsprozesses (Tokenisierung, Taggingverfahren, Tagset) ergeben;

(iii) ein Tagset, das für die Aufgabe des Wortartentaggings für Genres internetbasierter Kommunikation erweitert wurde;

(iv) Trainingsdatensets mit manuell annotierten Daten, denen die Kategorien des erweiterten Tagsets zugrunde liegen und die die unter Punkt (ii) formulierten Probleme in einer für linguistische Analysezwecke sinnvollen Art und Weise behandeln (Goldstandard);

(v) Verarbeitungswerkzeuge, die auf diesen Trainingsdaten auf die Anwendung des erweiterten Tagsets und auf den Umgang mit den unter (ii) beschriebenen Problemtypen trainiert wurden.

Der vorliegende Beitrag präsentiert Beschreibungen und Kategorien zu den Punkten (i) und (ii) und leitet daraus Vorschläge für die Erweiterung des STTS ab (Punkt iii). Er stellt damit die wesentlichen konzeptuellen Grundlagen bereit, um Trainingsdatensets (iv) aufzubauen und auf deren Basis Verarbeitungswerkzeuge auf den Umgang mit den sprachlichen Besonderheiten internetbasierter Kommunikation zu trainieren (v).

Der Beitrag ist wie folgt aufgebaut: Nach einer Vorbemerkung zum Status sprachlicher „Besonderheiten“ und zu deren Verteilung in Kontexten internetbasierter Kommunikation (Abschnitt 2) skizzieren wir in Abschnitt 3 eine Typologie charakteristischer Phänomene bei der schriftlichen Sprachverwendung in Genres internetbasierter Kommunikation. Die in der Typologie erfassten Phänomentypen werden anhand von Datenbeispielen aus Chats und aus Wikipedia-Diskussionsseiten veranschaulicht. Der Fokus der Typologie liegt auf sprachlichen Einheiten und auf Phänomenen der schriftlichen Realisierung. Einheiten, in denen sich sprachliche Ausdrücke bzw. Schriftlichkeitsphänomene mit Phänomenen hypermedialer Vernetzung überlagern (z.B. Hashtags in Tweets), bedürfen weiterer empirischer und konzeptueller Klärung und werden daher zwar erwähnt, aber nicht systematisch mitbehandelt (vgl. Abschnitt 3, Phänomentyp VII).

In Abschnitt 4 geben wir einen datengestützten Überblick über Probleme bei der Verarbeitung einiger der in Abschnitt 3 vorgestellten Phänomene mit existierenden Werkzeugen für die automatische Tokenisierung und Wortartenannotation deutscher Sprachdaten, die von linguistischen Anwendern „off the shelf“ über die Oberfläche der webbasierten Analyseplattform *WebLicht* (https://weblicht.sfs.uni-tuebingen.de/) genutzt werden können und die das „Stuttgart-Tübingen-Tagset“ (STTS) als Ressource für die Zuordnung von Wortartentags zu Wort-Tokens nutzen. Der Problemaufriss zeigt, dass die

Probleme bei der automatischen Zuordnung von Wortartenkategorien auf unterschiedlichen Ebenen des Verarbeitungsprozesses zu verorten sind. Betroffen ist nicht nur die Ebene des Wortartentaggings selbst, sondern auch die Ebene der Tokenisierung sowie die Ebene der Festlegung eines geeigneten Kategorien- und Tagsets. Entsprechend müssen Ansätze zur Optimierung des Wortartentaggings nicht nur Verarbeitungsverfahren auf verschiedenen Analyseebenen an die Besonderheiten der Domäne „Sprachverwendung in der internetbasierten Kommunikation" anpassen; vielmehr müssen auch die Tagsets, die diesen Verfahren als Ressource zugrunde liegen, an die sprachlichen Besonderheiten der Domäne angepasst werden. Abschnitt 4 zeigt, welche der in Abschnitt 3 vorgestellten Phänomene einer Anpassung der Verfahren und welche einer Erweiterung des POS-Tagsets bedürfen. Für Phänomene der letzteren Art formulieren wir in Abschnitt 5 einen Vorschlag, wie sie in einer Erweiterung des STTS berücksichtigt werden könnten.

2 Sprachliche Besonderheiten und sprachliche Variation in der internetbasierten Kommunikation

Die schriftliche Sprachverwendung in Genres internetbasierter Kommunikation weist eine Reihe von Phänomenen auf, hinsichtlich derer sie sich von der Sprachlichkeit in redigierten Texten unterscheidet. Als „Besonderheiten" erscheinen diese Phänomene unter zweierlei Perspektiven:

a) Im Vergleich mit redigierten Texten (z.B. journalistischen Genres), die eine hohe Konformität mit den Normen der geschriebenen Standardsprache aufweisen, erscheinen sie als Abweichungen von der Norm bzw. als sprachliche Mittel, die in standard-nahen Textsorten nicht oder nur in besonderen Fällen auftreten oder erwartbar sind;

b) aus der Perspektive der Automatischen Sprachverarbeitung erscheinen sie als Phänomene, die mit gängigen, an redigierten Texten trainierten Verarbeitungswerkzeugen nicht oder nur unzureichend analysiert werden können und die deshalb bei der linguistischen Annotation von Sprachdaten einer besonderen Behandlung bedürfen.

Die Etikettierung der schriftlichen Sprachverwendung in der internetbasierten Kommunikation als ‚abweichend von den schrift(sprach)lichen Normen' bzw. ‚nichtstandardisiert' ist dabei lediglich im Sinne einer ‚Nicht-*Bezogenheit* auf den schriftlichen Standard' aufzufassen und nicht als Abweichung im Sinne einer nur unzureichend umgesetzten oder defizitären Erfüllung der Anforderungen an standardschriftlich realisierte Texte. Im Gegensatz zu prototypischen Textäußerungen, die – etwa i.S.v. Ehlich (1983, 1984) – situationsunabhängig rezipiert und verstanden werden sollen, unterliegt die sprachliche Gestaltung von Beiträgen im Rahmen getippter Dialoge in Foren, Chats, Weblog-Kommentaren, Diskussionen in Wikis oder auf den Profilseiten sozialer Netzwerke eigenen Normen, die den Normen für die Gestaltung mündlicher Gesprächsbeiträge nicht unähnlich sind: Leitlinie bei der Gestaltung der Kommunikationsbeiträge ist weniger die Sicherung der situationsunabhängigen Verständlichkeit eines sprachlichen *Produkts* als vielmehr der kommunikative Erfolg der damit realisierten sprachlichen Handlung(en) im Kontext der laufenden Interaktion. Die

sprachliche Form wird dabei optimiert für Adressaten, die die im Kommunikationsgeschehen vorausgegangenen kommunikativen Schritte kennen und über den aktuellen Stand des Geschehens auf dem Laufenden sind (vgl. Storrer 2012, 2013).

Dies gilt nicht nur für synchrone Genres wie Chat und Instant Messaging, in denen alle Beteiligten zeitgleich auf das Kommunikationsgeschehen orientiert sind und die Vorbeiträge noch unmittelbar mental präsent haben, sondern auch für asynchrone dialogische Genres wie Diskussions-Threads in Online-Foren, Wikis und sozialen Netzwerken, bei denen die schriftlichen Beiträge entsprechend ihrer zeitlichen Abfolge und häufig auch thematisch strukturiert auf einer Bildschirmseite vorgehalten werden. Auch wenn die Beteiligten nicht zeitgleich auf die Fortentwicklung des Dialoggeschehens orientiert sind, ist hier die jeweilige Vorkommunikation jederzeit im Wortlaut nachlesbar und kann, da neue Beiträge am Bildschirm direkt an den dokumentierten Verlauf der Vorkommunikation angefügt werden, bei der Formulierung neuer Beiträge auch bei den Adressaten eine Orientiertheit über den aktuellen Stand des Kommunikationsgeschehens vorausgesetzt werden.

Die in Abschnitt 3 vorgestellte Typologie beschreibt sprachliche Besonderheiten, die *charakteristischerweise* in der internetbasierten Kommunikation auftreten – was nicht bedeutet, dass sie in jedem Genre, in jedem Nutzungskontext und in jedem konkreten Kommunikationsereignis in gleicher Frequenz und Verteilung vorzufinden sind. Vielmehr hat die neuere linguistische Forschung zur internetbasierten Kommunikation wiederholt nachgewiesen, dass die Nutzer internetbasierter Kommunikationstechnologien ihre Sprachverwendung in ähnlicher Weise an soziale, institutionelle, situative und individuelle Rahmenbedingungen anpassen sowie in Abhängigkeit zu Themen und zu den jeweils instanziierten kommunikativen Gattungen variieren, wie dies in mündlichen Gesprächen oder in schriftlichen Texten der Fall ist. So zeigen z.B. Androutsopoulos/Ziegler (2003), dass die Verwendung von Regionalismen in stadtspezifischen Chat-Kanälen nicht unabhängig von der Dialekt-Standard-Relation der zugehörigen „realweltlichen" Region zu denken ist. Jarbou/al-Share (2012) beschreiben in einer Korpusuntersuchung zu Schreibvarianten in jordanischen Chats die Abhängigkeit graphematischer Variation von Gender und Dialekt. Storrer (2013) weist in einer quantitativen Auswertung eines Korpus zur deutschen Wikipedia nach, dass das Auftreten „typischer" Stilelemente systematisch in Abhängigkeit zum Genre (Wikipedia-Artikelseiten vs. Wikipedia-Diskussionsseiten) variiert. Luckhardt (2009) arbeitet in ihrer Dissertation auf Basis der Analyse eines Chat-Korpus heraus, dass die Verwendung „typischer" Stilmerkmale der internetbasierten Kommunikation zudem auch stark von individuellen Präferenzen der Nutzer beeinflusst wird. Auch Arbeiten, die sich mit der Nutzung der Chat-Technologie für unterschiedliche professionelle Nutzungskontexte (Beratung, Bildung, Medien) beschäftigen, haben differenziert die technischen und sozialen Faktoren beschrieben, die auf die Struktur und sprachliche Gestaltung des kommunikativen Austauschs Einfluss nehmen und die für eine erfolgreiche Nutzbarmachung der Technologie in professionellen Kontexten zu kontrollieren sind (vgl. z.B. die Beiträge in Beißwenger/Storrer (Hrsg.) 2005 sowie das „Chat-Szenario"-Modell in Beißwenger/Storrer 2005a).

Internetbasierte Kommunikationstechnologien konstituieren Kommunikations*formen* und keine *kommunikativen Gattungen*. Die *Formen* (z.B. Chat-Kommunikation, Forenkommu-

nikation, Twitter-Kommunikation usw.) sind durch die technischen Rahmenbedingungen der Technologie determiniert; einzelne *Gattungen* werden hingegen erst in der Nutzung einer Form für konkrete kommunikative Zwecke instanziiert (vgl. Beißwenger 2003, 2007: 107–112 in Anlehnung u.a. an vergleichbare Differenzierungen für den Bereich der Textlinguistik in Brinker 2001; Dürscheid 2005a). Entsprechend gibt es nicht „*die* Sprache des Internet“ oder „*die* Sprache des Chat / der E-Mail / der sozialen Netzwerke / der Blogs“, die sich nur mit Bezug auf die technologischen Rahmenbedingungen und in Absehung von sozialen und pragmatischen Faktoren bestimmen ließe (Storrer 2000, Storrer 2009; Dürscheid 2004).

Die Zusammenstellung verschiedener existierender oder derzeit in Aufbau befindlicher Korpora zur internetbasierten Kommunikation trägt dieser Tatsache Rechnung und berücksichtigt systematisch verschiedene Kommunikationsformen und deren Nutzungskontexte – vgl. z.B. Reynaert et al. (2010) zur Zusammensetzung des niederländischen Referenzkorpus *SoNaR*, Beißwenger (2013) zum 2002-2008 aufgebauten, nach Handlungsbereichen differenzierten „Dortmunder Chat-Korpus“, Beißwenger et al. (2013) zum laufenden Projekt „Deutsches Referenzkorpus zur internetbasierten Kommunikation“ (*DeRiK*).

3 Für das Wortartentagging relevante sprachliche Besonderheiten: eine Typologie

In diesem Abschnitt skizzieren wir eine Typologie sprachlicher Besonderheiten in Genres internetbasierter Kommunikation. Die unterschiedliche Verteilung der in der Typologie beschriebenen Phänomene in unterschiedlichen kommunikativen Genres sowie unter unterschiedlichen sozialen, institutionellen, situativen und individuellen Rahmenbedingungen (vgl. Abschnitt 2) wird dabei ausgeblendet. Ziel der Typologie ist es nicht, sprachliche und stilistische Variation in der internetbasierten Kommunikation zu beschreiben, sondern eine differenzierte Übersicht über diejenigen Phänomene zu bieten, hinsichtlich derer sich die schriftliche Sprachverwendung in der internetbasierten Kommunikation von der Sprachverwendung in redigierten Texten unterscheidet, um auf dieser Grundlage Verfahren für die automatische Erkennung und Disambiguierung dieser Phänomene entwickeln zu können. Der Fokus liegt dabei auf Einheiten, die bei der automatischen Wortartenannotation (POS-Tagging) Probleme bereiten. Da POS-Annotationen die Basis für alle weiteren Ebenen der linguistischen Annotation von Korpora darstellen, ist diese Annotationsebene beim Aufbau von Korpora von zentraler Bedeutung. Erst die Behandlung der in der Typologie erfassten Phänomene sowie der zugehörigen Verarbeitungsprobleme mit Werkzeugen für die automatische Sprachanalyse (s. Abschnitt 4) wird es ermöglichen, große, reichhaltig linguistisch annotierte Korpora zur internetbasierten Kommunikation aufzubauen, mit denen die Verteilung IBK-typischer sprachlicher Phänomene relativ zu unterschiedlichen Kontexten auf breiter Datenbasis quantitativ und qualitativ untersucht werden kann.

Um die in der Typologie erfassten Phänomene und die zugehörigen Verarbeitungsprobleme in den Griff zu bekommen, bieten sich prinzipiell zwei Ansätze an:

– die *Normalisierung*, bei der Wortformen, die von den genutzten Verarbeitungswerkzeugen aufgrund nicht-standardkonformer Formmerkmale nicht

sinnvoll analysiert und klassifiziert werden können, in einem Vorverarbeitungsschritt auf einer separaten Annotationsebene in eine Form überführt werden, die diese Merkmale nicht mehr aufweist. Dabei werden z.B. Phänomene geschriebener Umgangssprache durch ihre standardkonformen Pendants ersetzt (*machste* ⇨ *machst du*, *wat isn* ⇨ *was ist denn* usw.) und Schnellschreibphänomene beseitigt (*idese aufgabe* ⇨ *diese Aufgabe*). Anschließend wird die normalisierte Version der Ausgangsdaten mit den Verarbeitungswerkzeugen analysiert und annotiert.

– die *Anpassung existierender oder die Entwicklung neuer, spezialisierter Verarbeitungsverfahren* für den Umgang mit den Probleme bereitenden Phänomenen.

Welcher Ansatz sich im Einzelfall als praktischer erweist, muss in Abhängigkeit von der Forschungsfrage sowie vom Status der Verarbeitungsprobleme bereitenden Phänomene im jeweiligen Projekt entschieden werden. Grundsätzlich kommt eine Normalisierung nur für solche Phänomene in Betracht, zu denen sich eine normalisierte Form angeben lässt (z.B. im Falle der kontraktierten Formen, die sich durch Aufhebung der Klise in zwei standardorthographisch reguläre Formen auflösen lassen: *machste* ⇨ *machst du*). Hingegen entziehen sich Einheiten wie Emoticons, Aktionswörter und Adressierungen, die als spezifisch für Genres internetbasierter Kommunikation gelten können, einer Normalisierung. Für ihre Behandlung bei der POS-Annotation ist eine Erweiterung des zugrunde liegenden Tagsets unumgänglich.

Die nachfolgende Typologie zielt darauf, Besonderheiten der schriftlichen internetbasierten Kommunikation möglichst feinkörnig zu erfassen und Phänomenbeschreibungen bereitzustellen, die als Grundlage für die Erweiterung von Tagsets, für die Erarbeitung von Normalisierungsverfahren und die Anpassung von Verarbeitungswerkzeugen dienen können.

Der Fokus der Typologie liegt auf *sprachlichen Besonderheiten* und auf *Phänomenen der schriftlichen Realisierung*. Einheiten, bei denen sich Sprachliches mit Phänomenen der hypermedialen Vernetzung überlagert, bedürfen weiterer empirischer und konzeptueller Klärung und sind nur am Rande berücksichtigt (Phänomentyp VII).

Die Experimente, die wir anschließend in Abschnitt 4 vorstellen, zeigen, welche der im Folgenden unterschiedenen Phänomene bei der automatischen Verarbeitung Probleme auf welchen Verarbeitungsebenen verursachen. Für diejenigen Phänomene, die nur durch Erweiterung der Tagsets in den Griff zu bekommen sind, formulieren wir in Abschnitt 5 einen Vorschlag, wie sie im STTS dargestellt werden könnten.

I. Schnellschreibphänomene

- I.1 Irreguläre Verwendung von Spatien
- I.2 Tippfehler
- I.3 Ökonomiebedingte Abweichung von den Normen für die Groß- und Kleinschreibung

II. Graphische Nachbildung suprasegmentaler Elemente der gesprochenen Sprache

- II.1 Vollgroßschreibung von Wortformen und ganzen Äußerungen
- II.2 Iteration von Graphemen
- II.3 Iteration von Interpunktemen

III. Geschriebene Umgangssprache

- III.1 Umgangssprachlich fundierte Wortschreibungen
- III.2 Umgangssprachliche kontraktierte Formen
- III.3 Umgangssprachliche Lexik

IV. Verfremdungsschreibungen

V. IBK-typische Akronyme

VI. IBK-spezifische interaktive Einheiten

- VI.1 Emoticons
- VI.2 Aktionswörter
- VI.3 Adressierungen

VII. Weitere Phänomene

Abb. 1: Typologie sprachlicher Besonderheiten in der internetbasierten Kommunikation.

Im Folgenden geben wir Kurzbeschreibungen zu den einzelnen Phänomentypen (und Subtypen) (Abb. 1) und erläutern diese anhand von Beispielen aus Wikipedia-Diskussionsseiten [WPD] und aus dem Dortmunder Chat-Korpus [CHAT]:

I. Schnellschreibphänomene

Da es den Kommunikationsbeteiligten mit ihren schriftlichen Beiträgen primär darum geht, die laufende Interaktion weiterzuentwickeln, und weniger darum, ein situationsunabhängig verständliches schriftliches Produkt herzustellen, geschieht die Planung und Versprachlichung von Beiträgen häufig schnell und spontan und auf eine Korrekturdurchsicht der Beiträge vor der Verschickung wird häufig verzichtet. Typische Begleiterscheinungen der schnellen Produktion sind Tippfehler, untypische Platzierungen von Spatien sowie ein Verzicht auf die Anwendung der Regeln für die Großschreibung am Wort- und Satzanfang:

I.1 Irreguläre Verwendung von Spatien

Beim schnellen Tippen werden Spatien entweder – bedingt durch den Anschlag der Space-Taste mit zu geringer Dynamik – ausgelassen (Beispiele 1 und 2) oder kommt es durch versehentliches Betätigen von Tasten in einer anderen als der anvisierten Abfolge zur Platzierung von Spatien an einer anderen als der anvisierten Stelle (Beispiel 3). Entsprechend ergeben sich im verschickten Beitrag Token-Grenzen, die nicht den Grenzen der vom Produzenten realisierten Wort-Tokens entsprechen:

(1) *nu is tombefreit* (anstelle von: *Tom befreit*) [CHAT]

(2) *sind die B-Städter jetzt virtuell on? odersollen wir warten?* [CHAT]

(3) **wünk* und wehe ihr verleumdet mich *grml* *brummel* imme rnoch nicht fassen kann* [CHAT]

Es ist anzunehmen, dass irregulär gesetzte Spatien in der Mehrzahl der Fälle unbeabsichtigt als Folge schneller Tippaktivität auftreten; in Einzelfällen können sie aber auch auf ein bewusstes Spiel mit der medialen Schriftlichkeit der Kommunikation zurückgeführt werden (vgl. das Beispiel in Beißwenger/Storrer 2012: 100).

I.2 Tippfehler

Tippfehler sind irreguläre orthographische Realisierungen von Wortformen, die typischerweise daraus resultieren, dass die Tasten der Tastatur – flüchtigkeitsbedingt – falsch, ungenau, zu stark, zu schwach oder in falscher Reihenfolge angeschlagen werden. Die in Genres internetbasierter Kommunikation auftretenden Tippfehler lassen sich in vier Typen einteilen (Beißwenger 2000: 73f.): „Vertipper" (Typ I), bei denen anstelle der Taste mit dem anvisierten Zeichen fälschlich eine auf der Tastatur nebengelegene Taste angeschlagen wird (Beispiel 4); „Vertipper" (Typ II), bei denen zusätzlich zur Taste mit dem anvisierten Zeichen eine weitere Taste angeschlagen wird, die entweder ein zusätzliches Zeichen realisiert oder – im Falle der Shift-Taste – eine Realisierung der in der Folge getippten Zeichen in Versalien verursacht (Beispiele 5 und 6); anschlagsdynamisch bedingte Auslassungen (Beispiel 7) oder Mehrfachrealisierungen eines anvisierten bzw. an der betreffenden Position in der Graphemstruktur vorgesehenen Zeichens (Beispiel 8); „Buchstabendreher", bei denen zwei oder mehrere in der Graphemstruktur aufeinander folgende Zeichen in falscher Reihenfolge realisiert werden (Beispiel 9):

(4) *Wer gib eigentlich seinen Namen hier preis? Ingeborbbachmann. Seltsam!* [CHAT]

(5) *31,5 werd ich auchg noch schaffen* [CHAT]

(6) *dass prinzipiell auchg alles, was sie "sAQGEN2; MITgeschnitten werden kann...* [CHAT]

(7) *Bin ja garnicht böe* [CHAT]

(8) *ich bin durcheinander weil ich nochmals von vorne anfangen muss du dagegenkannst anffangen zu planen* [CHAT]

(9) *da sollten wir wohl schleunigst den mantel des schweigens über idese aufgabe breiten / dat ebste findeste eigentlich wenn du gar nich suchst* [CHAT]

I.3 Ökonomiebedingte Abweichung von den Normen für die Groß- und Kleinschreibung

Als typisches Schnellschreibphänomen gilt weiterhin die liberale Anwendung oder radikale Nichtanwendung der Normen für die Großschreibung. Dabei wird, um Tippaufwand zu minimieren, auf die Betätigung der Umschalttaste verzichtet, die dazu benötigt wird, den Buchstaben, mit dem eine gleichzeitig angeschlagene Buchstabentaste auf der Tastatur belegt ist, in Versalien zu realisieren. Im Einzelfall kann die irreguläre Anwendung der Regeln für die Großschreibung im Deutschen natürlich auch kompetenzbedingt sein.

Das Abweichen von bzw. die Nichtanwendung der Großschreibung kann sich dabei sowohl auf die Großschreibung von Substantiven und Eigennamen wie auch auf die Großschreibung von Satzanfängen (innerhalb von Nutzerbeiträgen) und Textanfängen (bezogen auf die internetbasierte Kommunikation: die Großschreibung zu Beginn eines neuen Kommunikationsbeitrags) beziehen:

(10) Reguläre Großschreibung der Substantive, aber keine Großschreibung am Beginn des Beitrags: *ich habe mich heute bei Wikipedia angemeldet (nach Inspiration durch meinen Berufsschullehrer *zwinker*) und hoffe auf einen Ausbau des Portals* [WPD]

(11) Großschreibung zu Beginn des Beitrags, aber konsequente Kleinschreibung aller weiteren Substantive: *Dinge, die körperlich nicht passieren können oder kommunikation, die eigetnlich widersinnig wird, wir sofort akzeptiert und lediglich als philosophischer bruch empfunden.* [CHAT]

(12) Radikale Kleinschreibung (Substantive, Satzanfänge, Beginn des Beitrags): *immer wieder schön beim arbeiten gestört zu werden werden so kleinigkeiten wie ein externer link. entweder machst du die seite oder lässt mich einfach mal alles hier fertig machen und dann kannse von mir aus mal mal drüber guggn.* [WPD]

II. Graphische Nachbildung suprasegmentaler Elemente der gesprochenen Sprache

Ein weiterer Phänomentyp ist die Nachbildung suprasegmental-prosodischer Elemente der gesprochenen Sprache mit den Mitteln der Schrift. Typische Verfahren sind die Vollgroßschreibung von Wortformen und ganzen Sätzen oder Beiträgen sowie die Graphemiteration (vgl. Beißwenger 2000: 104f.). Dürscheid (2005: 48f.) klassifiziert Normabweichungen dieser Art als *Stilmittel*; im Gegensatz zur liberalen Anwendung der Normen für die Groß- und Kleinschreibung, die zwar – sofern es sich nicht um Kompetenzfehler handelt – ebenfalls beabsichtigt ist, aber nur der Ökonomisierung der Beitragsproduktion dient, handelt es sich bei den Phänomenen in dieser Gruppe um Schreibungen, mit denen zudem ein kommunikativer Zweck verfolgt wird:

II.1 Vollgroßschreibung von Wortformen und ganzen Äußerungen

Die Vollgroßschreibung von Wortformen, Sätzen oder ganzen Beiträgen nutzt Versalien als typographische Metapher für Intensität. Dabei werden Großbuchstaben verwendet, um lautes Sprechen oder Schreien darzustellen. Groß geschrieben werden entweder einzelne Wortformen oder ganze Sätze bzw. Beiträge. In ersterem Fall (Beispiele 13–16) dient die Vollgroßschreibung häufig der Fokussierung, in zweiterem Fall der Symbolisierung von Emphase bzw. von lautem Sprechen/Schreien (Beispiel 17):

(13) *Mehr Arbeitsplätze müssen her, die Konjunktur muß angekurbelt werden. Rot/Grün setzt die FALSCHEN Signale, trifft die FALSCHEN Entscheidungen .....armes Deutschland .. zahlen müssen leider jetzt auch die, die Rot/Grün nicht gewählt haben.. :-(((* [CHAT]

(14) *latinumsprüfung ist NICHT egal... aber ich nehme mal an, du meintest das auch nicht...* [CHAT]

(15) *Hab übrigens mitm Hersteller Kontakt aufgenommen, die wollen das Logo erst dann hochladen, wenn der Löschantrag NICHT durchgegangen ist* [WPD]

(16) (Autor 1:) *Einfach die Tabelle aus einer alten Version in die Zwischenablage kopieren, in die letzte Version einfügen, speichern, fertig :-)* – (Autor 2:) *:-) DAS habe ich ja die ganze Zeit versucht! *heul* Und ich habe es an zwei Computern mit drei verschiedenen Browsern probiert: Nullanzeige :-(* [WPD]

(17) *mathe mündlich? MATHE MÜNDLICH! BRUTAL! das war bei uns schriftlich und schon schlimm genug!* [CHAT]

II.2 Iteration von Graphemen

Die Iteration von Graphemen dient der Nachbildung von Dehnung zur Setzung von Emphaseakzenten in der gesprochenen Sprache und zum Ausdruck von Emotion:[3]

(18) *ein seeeehr heikles Thema auf jeden Fall, wer da einen fairen und treffenden Absatz zustande bringt, bekommt von mir einen Orden;-)* [WPD]

(19) *Tiggi ist ja soooooo erwachsen ;-)* [CHAT]

(20) *leider nicht, nö *schaaaaade* [CHAT]

Die Iteration tritt auch in akronymischen, silbisch interpretierten Formen auf (im folgenden Beispiel im Aktionswort *lol*):

(21) *FRagt mal den Koch in Hessen nach dem Spendenkonto*looooooool** [CHAT]

Bisweilen treten Vollgroßschreibung (II.1) und Graphemiteration auch in Kombination auf:

(22) *Burzel nimmt anlauf und stürzt sich im HUUUUUURRRRRAAAAAA auf SX zum SUPERDUPPERHYPERMEGAKKKKKNNNNUUUUUDDDDLLLLÄÄRRRRR* [CHAT]

II.3 Iteration von Interpunktemen

Die Iteration von Interpunktemen dient der Symbolisierung von Emphase und Emotion (Beispiele 23 und 24) bzw. der (performativen) Nachbildung von Planungspausen aus der gesprochenen Sprache im Medium der Schrift (Beispiel 25):

(23) *Wir wollen die Welt verbessern... OHNE Drogen!!!!!* [CHAT]

(24) *Jeder muss zugeben: dass kanns doch nicht sein!!!* [WPD]

(25) *Warum stehe ich nur mit Klarnamen in dieser Kategorie?.....mmh..... Werbung?.....Schwarzes Brett???.....* [WPD]

Während die Vervielfältigung von Vokal- und Konsonantengraphemen eine Iteration der graphischen Repräsentationen einzelner Lautsegmente *innerhalb* eines Tokens darstellt, wird bei der Iteration eines Interpunktionszeichens das komplette Token (das standardgemäß nur aus einem einzigen Zeichen besteht) vervielfältigt.

III. Geschriebene Umgangssprache

Mit der internetbasierten Kommunikation ist die Schrift als Realisierungsmedium für sprachliche Äußerungen im großen Stil auch für solche Handlungsbereiche nutzbar geworden, die zuvor eher der gesprochenen Sprache vorbehalten waren (vgl. Storrer 2001:439). Insbesondere in solchen Nutzungskontexten internetbasierter Kommunikationstechnologien, die neben der Dialogizität weitere Kommunikationsbedingungen der Nähe – z.B. einen vertrauten, informellen Umgang der Partner, eine freie Themenentwicklung, eine hohe Spontaneität bei der Planung und Realisierung von Kommunikationsbeiträgen – aufweisen, lässt sich eine Orientierung an einer „Sprache der Nähe“ (i.S.v. Koch/Oesterreicher 1994) feststellen, die sich auf Wortebene in einer Realisierung umgangssprachlicher (z.T. auch mundartlicher) Formen und Strukturen im Medium der Schrift niederschlägt („Geschriebene Umgangssprache“, vgl. z.B. Kilian 2001).

III.1 An der umgangssprachlichen Lautung orientierte Wortschreibungen

Bei der Wortschreibung zeigt sich die Orientierung an der Umgangssprache in der Anwendung der orthographischen Prinzipien des Deutschen auf die umgangssprachliche anstelle der standardsprachlichen Lautung von Wortformen. Dabei werden z.T. auch typisch sprechsprachliche Elisionen ins schriftliche Medium transponiert (wie z.B. *is* < *ist*, *ne* < *eine*, *wunder* < *wundere* in den nachfolgenden Beispielen):

(26) *Jut, ich find die Variante mit "Die" auch besser und "richtiger" in diesem Fall.* [WPD]

(27) *japp da habe ich es auch gelesen, aber sonst nirgens* [WPD]

(28) *Das sind zuviele Artikel drinne, die mit Recht im weiteren Sinne nichts mehr zu tun haben.* [WPD]

(29) *Nee, mit Vadder is hier Kim Il Sung gemeint, der Sohn demnach Kim Jong Il.* [WPD]

(30) *Isch ja gut, es hier noch anderes zu tun als solcher Kleckerleskram.* [WPD]

(31) *ick wunder mir über jarnischt mehr* [WPD]

(32) *Ick weeß, auch det is Jeschichte ... aber hässlich bleibt hässlich.* [WPD]

(33) *Noch ne kleene Nachfrage: Brauchen wir demnächst ooch noch ne Themenkat:NS-Opfer?* [WPD]

(34) *entweder machst du die seite oder lässt mich einfach mal alles hier fertig machen und dann kannse von mir aus mal mal drüber guggn.* [WPD]

(35) *Kandidaturdiskussion auf Artikeldiskussion kopieren und auf WP:KALP löschen. Dann noch einen Abbruchvermerk, Kandidaturbaustein aus Artikel entfernen - zack, feddich!* [WPD]

(36) *Moinsen, wirf mal einen Blick auf WP:WEB, das hatte ich dir schon heude mittach als IP geraten* [WPD]

(37) *dit is selbstverständlich* [CHAT]

(38) *ozelot sacht moin zu stoeps* [CHAT]

(39) *p.s.: knuffine ist ne olle petze ;) *lol** [CHAT]

(40) *gun tach* [CHAT]

III.2 Umgangssprachliche kontraktierte Formen

In der internetbasierten Kommunikation finden sich unterschiedlichste Arten von kontraktierten Formen, die aus der Verschmelzung zweier aufeinanderfolgender syntaktischer Wörter resultieren. Einerseits begegnen kontraktierte Formen wie *im* (< *in dem*), *am* (< *an dem*), *zur* (< *zu der*), *zum* (< *zu dem*), *ans* (< *an das*), *ins* (< *in das*), die auch standardsprachlich – d.h. in redigierten Texten – gebräuchlich sind, einen hohen Grammatikalisierungsgrad aufweisen und sich in vielen Fällen nicht mehr ohne Weiteres durch ihre Ausgangsformen ersetzen lassen (z.B. in *zum Erliegen/Stocken/Erlahmen kommen*, *ins Schleudern/Trudeln/Grübeln/ Schwärmen geraten*). Daneben finden sich aber auch kontraktierte Formen wie *haste*, *biste*, *isn* und *aufm*, bei denen zwei (oder mehrere) aufeinanderfolgende Wortformen koartikulationsbedingt miteinander

verschmolzen werden und die somit als typisch sprechsprachlich zu gelten haben. Mit ihrer Übernahme in geschriebene Äußerungen – bei deren Produktion Koartikulation keine Rolle spielen *kann* – wird ein Phänomen medialer Mündlichkeit in die mediale Schriftlichkeit transponiert (‚verschriftet' i.S.v. Koch/Oesterreicher).

Typische Bildungsmuster sind (z.B.):

– Präposition + Artikel:	*innem, aufm, aus(s)n*
– Adverb + Artikel:	*noch(e)n*
– Konjunktion + Personalpronomen:	*fallste* (< *falls du*), *obse* (< *ob sie*)
– Auxiliarverb + Personalpronomen:	*haste, biste*
– Vollverb + Personalpronomen:	*machste, gehste, denkste, schreibste*
– Vollverb + zwei Personalpronomina:	*machstes, gibstes*
– Kopulaverb + Personalpronomen:	*warens*
– Modalverb + Personalpronomen:	*kannste, willste, sollste, darfste*
– Auxiliarverb + Abtönungspartikel:	*(was) isn (passiert)*

Beispiele für die Verwendung in Wikipedia-Diskussionen und in Chats sind:

(41) *Prinzipiell schon, wobei auch das mit überregional so ne Sache ist. Innem halbwegs großen Staat erreichen seriöse Regionalzeitungen deutlich mehr Leser als die gesamte Einwohnerschaft Kosovos.* [WPD]

(42) *wieso "war" aswad denn eine band? sind die nich dieses jahr aufm summerjam, d.h. noch oder wieder aktiv?* [WPD]

(43) *Ich hab's auch nicht verstanden, rinn' inne Kartoffeln, raus aussn Kartoffeln. Wenn's der Account so will...* [WPD]

(44) *Nochen Vorschlag: Die Episodenliste unter „Die Simpsons: Episodenliste" ablegen.* [WPD]

(45) *nicht so gut wie alt biste und wo kommste her* [CHAT]

(46) *Fallste das kannst, ist dir wohl der Nobelpreis sicher.* [WPD]

(47) *Kategorie:Gesellschaft ist schlicht falsch; obs Dir passt oder nicht.* [WPD]

(48) *...war das die ursprüngliche zielsetzung? oder wars nich einfach nur der wunsch, neulingen bei ihren erstschritten zu helfen - unabhängich davon, obse auch bleibn... ??* [WPD]

(49) *hm, magste mal genauer erläutern, warum ein exzellenter artikel über eine skisprungschanze „fatal" sei?* [WPD]

(50) *alles klar, ich schreibs nochmal neu* [WPD]

(51) *na da haste aber was verschlimmbessert, machstes selber rückgängig? :)* [WPD]

(52) *naja,bei mir warens einige semester nachrichtentechnik halt.....* [CHAT]

(53) *aber wie gesagt,meinerseits hättestes ok da du ja nun quasi angefragt hast...* [CHAT]

(54) *Aber meine Überleitung zu den einzelnen Strömungen des Renaissancehumanismus sollste nicht wegkürzen* [WPD]

(55) *warum wollteste jemanden trffen der so etwas macht* [CHAT]

(56) *Könnteste die Datei nochmal auf Commons hochladen und hier SLA stellen? Danke.* [WPD]

Im Einzelfall können sich in ein- und demselben Token unterschiedliche Phänomene überlagern. Die Beispiele 57 und 58 zeigen umgangssprachliche kontraktierte Formen, die zusätzlich einen Tippfehler (57) bzw. eine Orientierung an einer umgangssprachlichen Formvariante (58) aufweisen:

(57) *dann kansstes dur auch direkt ausdenken...* [CHAT] (kannstes < kannst du es)

(58) *gibbet denn da kein bild?* [CHAT] (gibbet < gibt et „gibt es")

III.3 Umgangssprachliche Lexik

Geschriebene Umgangssprache findet sich nicht nur in der Graphie (III.1) und in der Verwendung kontraktierter Formen (III.2), sondern auch im Bereich der Lexik. Charakteristisch insbesondere für informelle Kontexte internetbasierter Kommunikation ist die Verwendung umgangssprachlicher, in der gesprochenen Sprache z.T. regional gebundener Lexik:

(59) *ozelot sacht moin zu stoeps* [CHAT]

(60) *Moinsen, wirf mal einen Blick auf WP:WEB, das hatte ich dir schon heude mittach als IP geraten* [WPD]

(61) *Und wer schreibt nun den Kurierbeitrag? Gruß vonne Maloche* [WPD]

(62) *p.s.: knuffine ist ne olle petze ;) *lol** [CHAT]

IV. Verfremdungsschreibungen

Nicht-standardkonforme Schreibungen wie in den Beispielen 63–65 lassen sich keinem der bislang unterschiedenen Phänomentypen zuordnen: Im Unterschied zu Schnellschreibphänomenen (I) lässt sich ihre Hervorbringung nicht aus der beschleunigten Textproduktion erklären, im Gegensatz zu Fällen des Typs II geht es bei ihnen auch nicht um die Nachbildung prosodischer Elemente. Gegenüber Fällen geschriebener Umgangssprache (III) wird gezielt vom phonographischen Prinzip (den Graphem-Phonem-Korrespondenzen für das Deutsche) abgewichen und für bestimmte Lautsegmente eine graphematische Repräsentation gewählt, die gegenüber der orthographischen Norm als stark markiert gelten kann. Funktion solcher Schreibungen

ist die gezielte und kreative graphematische Verfremdung – z.T. in Anlehnung an den Schreibgebrauch in bestimmten Szenen (z.B. graphematische Ersetzung <s> → <z> im Auslaut in Anlehnung an Schreibungen in der Hip-Hop-Szene, Beispiele 63 und 64). Auch die Nutzung von Zahlzeichen zur Repräsentation von Wörtern oder Wortbestandteilen, die phonologisch der Lautung des Zahlzeichens entsprechen (*n8* < *Nacht*, Beispiel 65), rechnen wir zu dieser Kategorie:

(63) *Congratulations Mädelz, das habt ihr gut gemacht. :)* [WPD]

(64) *Tach Wurm, geh mich doch fott mit die Plörre...Gibt's heute übrignz das nächste Waterloo, und ganz ohne Bierbecherweitwurf ???* [WPD]

(65) serIan: *ich hau mich ne ecke aufs ohr....*
stoeps: *n8 seri*
Tigerelse: *nacht, seri :)* [CHAT]

V. IBK-typische Akronyme

In Genres internetbasierter Kommunikation finden sich zum einen okkasionelle Abkürzungen, bei denen um der Zeitersparnis willen Ausdrücke für als bekannt vorausgesetzte Redegegenstände mit den Mitteln der deutschen Kurzwortbildung regulär gekürzt werden. Daneben haben sich für bestimmte, häufig genutzte Wendungen stabile Akronyme eingespielt, die einen z.T. hohen Idiomatisierungsgrad aufweisen. Viele der Abkürzungen und Akronyme stammen aus dem Englischen (Beispiele 68–76; deutschen Ursprungs hingegen 64–66 sowie *lg* in 67), einige sind spezifisch für bestimmte Communities (z.B. Wikipedia-Diskussionen: *POV* in Beispiel 76), andere sind Community-übergreifend gebräuchlich.

(64) *Einfach kann jeder, fragen kostet nix. Vllt. kann Dir Leonhardt ja schon 'ne Lageinfo liefern?* [WPD] (Vllt. < vielleicht)

(65) *Positionen zu Umweltpolitik. Jmd fleißig genug, die zu finden und einzuarbeiten?* [WPD] (Jmd < jemand)

(66) *kA was das kostet* [CHAT] (kA < keine Ahnung)

(67) *hi FIST, du was muss ich tun wenn ich eine neue Kategorie anlegen will bzw. weißt du wo die Seite ist wo man das nachlesen kann? Mir ist es dabei wichtig zu wissen welche Kriterien erfüllt werden müssen. thx und lg* [WPD] (thx < thanks, lg < liebe Grüße)

(68) *Dass die Veikkausliiga im Sommer spielt, führt hier IMO zu weit, das gehört in den entsprechenden Artikel* [WPD] (IMO < in my opinion)

(69) *Imho ist die Kritik an der Interpretation nicht richtig.* [WPD]
(Imho < in my humble opinion)

(70) *Btw: Diesen Edit solltest du genauer erklären.* [WPD] (Btw < by the way)

(71) *Sry, keine Ahnung warum der Artikel in der Form überhaupt existiert, für mich eindeutig enz. irrelevant, und SLA-fähig* [WPD] (sry < sorry)

(72) *Das Reisen mit Buchungsbestätigung statt klassischem Papierticket ist ja z.B. inzwischen afaik heutzutage eher die Regel als eine Ausnahme.* [WPD] (afaik < as far as I know)

(73) *Danke für den Hinweis - wird ASAP geändert.* [CHAT] (ASAP < as soon as possible)

(74) *cu biene bis samstag* [CHAT] (cu < see you)

(75) *bin mal eben für ca. 5 minuten "afk" - melde mich dann gleich wieder zurück.* [CHAT] (afk < away from keyboard)

(76) *Was ist daran so schwer zu verstehen, dass die Aussage „xy schmeckt gut" POV ist?* [WPD] (POV < point of view)

VI. IBK-spezifische interaktive Einheiten

Auf lexikalischer Ebene finden sich mit den *Emoticons*, den *Aktionswörtern* und den *Adressierungen* Elemente, die als spezifische Erweiterungen des Inventars sprachlicher Einheiten in der internetbasierten Kommunikation gelten können. Allen drei Elementen ist gemeinsam, dass sie sich syntaktisch-positional wie auch hinsichtlich ihrer Funktionen sehr ähnlich verhalten wie Einheiten, die in Grammatiken des Deutschen – in z.T. unterschiedlichem kategorialen Zuschnitt – als *Interjektionen* (GDS, Duden-4[7]), *Responsive* (GDS) oder *Gesprächspartikeln* (Duden-4[5]), in Grammatiken des Englischen als *interjections* (Greenbaum 1996, McArthur et al. 1998, Blake 2008), *inserts* (Biber et al. 1999; 2002) oder *discourse markers* (Schiffrin 1986) beschrieben werden: Sie sind in aller Regel nicht syntaktisch integriert, tragen also nicht zum kompositionalen Aufbau der Satzbedeutung bei, und können sowohl im linken oder rechten Außenfeld von Sätzen auftreten wie auch an nahezu beliebiger Position in Form von Parenthesen eingeschoben werden. Funktional sind sie spezialisiert auf Aufgaben im Bereich der Handlungskoordination im Dialog, der emotionalen Kommentierung und der Respondierung vorangegangener Partneräußerungen.

Die Grammatik der deutschen Sprache (*GDS*, Zifonun et al. 1997) führt Interjektionen (*ach*, *äh*, *mhm*, *ne*, *tja*) und Responsive (*ja*, *nein*, *okay*) aufgrund ihrer auf die spezifischen, auf die Organisation interaktiven Austauschs spezialisierten Funktionen in einer eigenen Kategorie ‚Interaktive Einheiten'. Will man Emoticons, Aktionswörter und Adressierungen in einen grammatischen Beschreibungsrahmen einordnen, so eignet sich diese Kategorie in besonderer Weise: Emoticons, Aktionswörter und Adressierungen lassen sich dann als spezifische Erweiterung des sprachlichen Inventars für die besonderen Bedürfnisse bei der Organisation *schriftlicher* dialogischer Interaktion beschreiben (vgl. Beißwenger et al. 2012: 3.5.1).

Als IBK-spezifische interaktive Einheiten sind Emoticons, Aktionswörter und Adressierungen auf unterschiedliche Aufgaben spezialisiert. Entsprechend bilden wir für ihre typologische Einordnung drei eigene Subtypen. Diesen sind die „klassischen“ Typen von interaktiven Einheiten – Interjektionen und Responsive – nebengeordnet (Abb. 2).

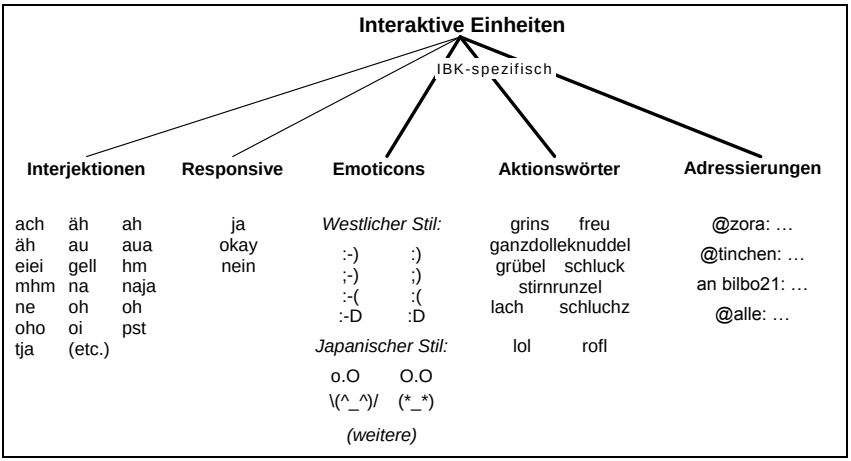

Abb. 2: Emoticons, Aktionswörter und Adressierungen als IBK-spezifische Erweiterungen der Kategorie der ‚interaktiven Einheiten‘ (GDS).

VI.1 Emoticons

Emoticons werden typischerweise durch die Kombination von Interpunktions- und Sonderzeichen gebildet; bisweilen können sie auch Buchstabenzeichen enthalten. Durch ihre ikonische Fundierung sind sie übereinzelsprachlich verwendbar. In unterschiedlichen Kulturkreisen haben sich unterschiedliche Stile herausgebildet (z.B. westlicher, japanischer, koreanischer Stil), deren Verwendung aber nicht auf die jeweiligen Ursprungskulturen beschränkt geblieben ist. So sind in vielen deutschsprachigen Online-Communities neben den „klassischen“ Emoticons westlichen Stils inzwischen u.a. auch japanische Emoticons gebräuchlich.

Emoticons können am Ende eines Satzes bzw. einer satzwertigen kommunikativen Einheit, am Ende eines Beitrags oder – seltener – in Form von Parenthesen auftreten sowie alleine einen Kommunikationsbeitrag realisieren. Sie werden u.a. zur emotionalen Kommentierung, zur Respondierung von Vorgängeräußerungen oder als Illokutions- und Ironiemarker verwendet:

(77) *och,die fischbude am heumarkt is ok**:-)*** [CHAT]
(Emotionale Kommentierung der eigenen Äußerung)

(78) ***:(((*** *Mit mir will einfach keiner chatten!* ***:(((*** [CHAT]
(Emotionale Kommentierung der eigenen Äußerung)

(79) *Ach nee, jetze isses plötzlich wieder eine Stadt?* ***:-P*** (WPD)
(Ironiemarkierung)

(80) ***:-/*** *Nein, nicht wirklich. Na ja, aber was ist den der Sinn des ganzen?0* [WPD] (Evaluative Respondierung einer Vorgängeräußerung)

(81) Weswolf: *Weswolf trabt zu tränchen rüber und begrüßt sie lieb*
tränchen: *tränchen hallöchen an alle :) besonders an WESWOLF :))))))))*
Weswolf: ***:-)))***

[CHAT] (Evaluative Respondierung einer Vorgängeräußerung)

(82) Raebchen: *ich habe als kind mal auf einem delphin gestanden (wirklich)*
Matrose: *cool*
Raebchen: *naja, der war gestrandet :(*
Matrose: *raebchen: und du bist dann darauf rumgeklettert - toll* ***:-(***

[CHAT] (Illokutionsmarkierung: Indikator für indirekten Sprechakt)

In Beißwenger et al. (2012: 3.5.1.4) wird für die Behandlung von Graphiken, die ähnliche Funktionen wie Emoticons und Aktionswörter übernehmen, eine Kategorie ‚interaction template' vorgeschlagen. Da wir uns in der hier beschriebenen Typologie auf (sprachliche bzw. tastaturschriftliche) Einheiten konzentrieren, die für das Wortartentagging relevant sind, sind graphisch realisierte Einheiten hier nicht als ein eigener Typ erfasst. Selbstverständlich ist bei der automatischen linguistischen Analyse von IBK-Daten aber auch mit in die Nutzerbeiträge eingebetteten Medienobjekten (Graphiken, animierte Graphiken, Videodateien) umzugehen – worunter auch die sog. „Graphik-Smileys" fallen –, dann aber nicht auf der Ebene der Wortartenklassifikation, sondern im Rahmen von Vorverarbeitungsschritten (vgl. VII zu weiteren Phänomenen in IBK-Daten). Entsprechend umfasst die hier beschriebene Kategorie ‚Emoticon' tatsächlich nur tastaturschriftlich erzeugte Einheiten.

VI.2 Aktionswörter

Aktionswörter sind einzelsprachlich gebundene symbolische Einheiten. Strukturell basieren sie auf einem Wort – zumeist einem unflektierten Verbstamm (‚Inflektiv', Schlobinski 2001) –, das entweder alleine steht oder um weitere Einheiten erweitert sein kann – im Falle von Inflektiven um vom Verb geforderte Ergänzungen oder Angaben. Im Falle solcher Konstruktionen werden die einzelnen Wortformen sehr häufig zusammengeschrieben, sodass sie formal

als ein Token erscheinen; sehr verbreitet ist zudem die Markierung mit ein- und ausleitenden Asterisken.

Aktionswörter werden zur Beschreibung von Gesten, mentalen Zuständen oder Handlungen verwendet. Sie dienen als Emotions- oder Illokutionsmarker (Beiträge 865, 876, 880 in Beispiel 83), als Ironiemarker (Beiträge 875, 878, 879), zur spielerischen Nachbildung fiktiver Handlungen (Beitrag 864) oder dazu, sich selbst (oder dem eigenen virtuellen Charakter) Charaktermerkmale oder innere Zustände zuzuschreiben (Beitrag 881).

Einige sehr gebräuchliche Aktionswörter haben die Form von Akronymen – z.B. **lol** (< *lauging out loud*), **rofl** (< *rolling on the floor laughing*), **g** (< *grin(s)*), **s** (< *smile*).

(83) Ausschnitt aus einem Chat-Mitschnitt:

858	Turnschuh:	*OHNE DEUTSCHLAND FAHRN WIR ZUR EM!*
859	system:	*Ryo hat die Farbe gewechselt*
860	Gangrulez:	*jo schade*
861	system:	*Windy123 geht in einen anderen Raum: Forum*
862	juliana:	*alle leute müssen ihre fernseher bei media markt bezahlen*
863	juliana:	*haha*
864	Turnschuh:	*Es gab mal ein Rudi Völler.....es gab mal ein Rudi Völler.....♫**sing**♫*
865	Ryo:	***g***
866	Gangrulez:	*hehe..das wurd eh gerichtlich gestoppt juliana*
867	juliana:	*echt?*
868	oz:	*gang: echt ??*
869	Gangrulez:	*ja*
870	juliana:	*wieso?*
871	Gangrulez:	*wettbewerbsverzerrung*
872	Naturkonstantler:	*Fussball ist sooo unendlich unwichtig...*
873	juliana:	*versteh ich nicht. ich fand es war ein cooler trick*
874	Gangrulez:	*aber es war eine Art Glücksspiel*
875	Turnschuh:	*mag auch keinen Fussball.....nur wollte ich das letzte Deutschlandspiel sehen **fg***
876	Chris-Redfield:	***s*** *aber net erlaubt @ juli*
877	juliana:	*fußball ist nen dreck wichtig. es ist ein spiel. hauptsache, die jungen männer haben sich fitgehalten und ihrer gesundheit was getan :)*
878	Gangrulez:	*und das entspircht nicht dem Handel **g***
879	juliana:	*chris, du weißt doch, daß ich ein gesetzesbrecher bin **g***
880	Chris-Redfield:	*ja ich weiß **s***
881	juliana:	***wildsei***

[CHAT]

VI.3 Adressierungen

Adressierungen sind sprachliche Ausdrücke, mit denen ein Kommunikationsbeitrag an einen bestimmten anderen Kommunikationsbeteiligten oder eine Gruppe von Kommunikationsbeteiligten adressiert wird. Ihr zentraler Bestandteil ist ein Ausdruck, mit dem der Adressat benannt (Angabe des Adressatennamens oder einer Variante davon) oder charakterisiert wird (Paraphrase). In vielen Fällen ist der Angabe des Adressaten ein Adressierungsmarker – häufig das auch aus E-Mail-Adressen bekannte <@>-Zeichen – vorangestellt. Sie stehen in aller Regel initial oder final zu der sprachlichen Äußerung, deren Adressiertheit kenntlich gemacht werden soll. Im Falle initialer Verwendung ist die Adressierung zumeist durch einen Doppelpunkt von den folgenden Einheiten abgesetzt.

Adressierungen dienen als ökonomische Mittel zur Anknüpfung an Beiträge oder Themen aus der Vorkommunikation oder zur Adressierung von Personen bzw. Personengruppen. In letzterer Funktion weisen sie Parallelen zu Anredeformen in Gesprächen auf; in ersterer Funktion dienen sie der Explizitmachung sequenzieller oder thematischer Kontinuität quer zur am Bildschirm angezeigten Beitragsabfolge. Da insbesondere in synchronen Formen internetbasierter Kommunikation (Chat, Instant Messaging) für die Beitragsproduzenten aufgrund der technischen Rahmenbedingungen nicht exakt planbar ist, ob handlungssequenziell oder thematisch aufeinander bezogene Beiträge am Bildschirm auch unmittelbar adjazent angezeigt werden, werden Adressierungen dazu eingesetzt, eine Rekonstruierbarkeit dieser Bezüge durch die Adressaten zu ermöglichen.

(84) Ausschnitt aus einem Chat-Mitschnitt (gekürzt):

4 Bates: *wir müssen aus baden raus, jasmin*
5 Jasmin: *weiß net..aber so weit ist stuttgart ja nicht weg*
[…]
8 Jasmin: *ja und dann auch noch nach schwaben..das ist eigentlich ne große überwindung für einen urbadner *g**
[…]
18 Bates: *ob die uns leben lassen?*
[…]
24 baloo: *loool@bates ich als franke trau mich auch nüber und hab es überlebt*
[…]
28 Biene: *baloo: ihr franken seid jaeh so ein völkchen *grins* nicht bös gemein, mein freund ist ja auch einer*
29 Jasmin: *@biene *gg**

Die Beiträge in Beispiel 84, das einem Dokument aus dem Dortmunder Chat-Korpus entnommen ist, enthalten drei Belege für Adressierungen. In Beitrag 24 ist die Adressierung an das Aktionswort *loool* angehängt; dadurch wird explizit markiert, dass dieser Teil des Beitrags der evaluativen Respondierung eines Vorbeitrags der Chatterin *bates* (Beitrag 18) dient; der folgende Teil des Beitrags (*ich als franke trau mich auch nüber und hab es überlebt*) knüpft daran dann thematisch an und entwickelt das aktuell verhandelte Thema weiter. In den Beiträgen 28 und 29 finden sich Adressierungen, die dem gesamten Beitrag vorangestellt sind; in 29 unter Verwendung eines Adressierungsmarkers, in 28 ohne Marker, dafür mit anschließendem Doppelpunkt, um den Ausdruck als von der syntaktischen Struktur der folgenden Einheiten abgesetzt zu kennzeichnen.

VII. Weitere Phänomene

Die Typologie erfasst unter den Subtypen I–VI ausdrücklich rein *sprachliche* Phänomene, d.h. solche Einheiten, die von den Kommunikationsbeteiligten per Tastatureingabe erzeugt werden. Nicht berücksichtigt sind „Graphik-Smileys“, die von den Nutzern per Auswahl aus einem Menü in ihre Beiträge eingefügt werden oder die vom System auf Basis bestimmter Tastatureingaben automatisch erzeugt werden. Solche – nicht-sprachlichen – Einheiten wären eher im Rahmen eines Vorverarbeitungsschrittes als auf der Ebene der Wortartenannotation zu behandeln; entsprechend sind sie in der hier beschriebenen Typologie nicht berücksichtigt (vgl. unsere Ausführungen zu Phänomentyp VI.1).

In der Typologie ebenfalls nicht berücksichtigt sind Einheiten, bei denen sprachliche Einheiten – häufig von den Autoren gezielt eingesetzt – mit Einheiten der Hypertextstruktur konvergieren. Beispiele für solche Einheiten sind Hashtags sowie Adressierungen in Tweets und in Facebook-Pinnwandeinträgen, die serverseitig automatisch zu Hyperlinks umgewandelt und von den Nutzern gezielt für die thematische Vernetzung ihrer Tweets mit Beiträgen anderer Nutzer (Hashtags in Twitter) bzw. für die Sichtbarmachung ihrer Beiträge auf den individuellen Startseiten der spezifizierten Adressaten (Adressierungen in Twitter und Facebook) genutzt werden. Die Beschreibung solcher Einheiten spielt für die Analyse von Kommunikationsbeiträgen in Genres internetbasierter Kommunikation zweifelsohne eine wichtige Rolle; um die Besonderheiten hypertextuell vernetzten Kommunizierens in Korpora zu beschreiben, können diese Einheiten aber nicht auf lediglich eine ihrer verschiedenen Funktionen – z.B. die sprachliche, die technische, die hypertextuelle – reduziert werden. Vielmehr gewinnen sie ihre spezifische Funktion im Rahmen der Kommunikation gerade durch die Konvergenz sprachlicher, technischer und struktureller Eigenschaften. In Beispiel 85 ist beispielsweise das Token „cornelsenverlag“ zugleich (i) ein Eigenname (in der „realen Welt“), (ii) der Name eines anderen Twitter-Nutzers (vermutlich der Twitter-Dependance des betreffenden Verlags), (iii) ein durch <@> gekennzeichneter Adressierungsausdruck, mit welchem die Adressierung des Tweets an den genannten Nutzer angezeigt wird, und (iv) ein Hyperlink, der dem Ausdruck vom System aufgrund der Verbindung mit dem <@>-

Zeichen automatisch hinzugefügt wird und der auf die Profilseite des adressierten Nutzers verweist. Durch die Kombination mit dem <@>-Zeichen fungiert der gesamte Ausdruck <@cornelsenverlag> darüber hinaus zugleich (v) als ein Kommando an das System, den Tweet in die persönliche Timeline des Nutzers „cornelsenverlag" zu integrieren; loggt sich der Nutzer das nächste Mal in Twitter ein, findet er den Tweet dort vor.

(85) Tweet-Nachricht auf twitter.com:

*Erfreulich, dass ich Vertreter der **@cornelsenverlag** e auf Veranstaltungen wie dem **#slml13** der **@werkstatt_bpb** antreffe. Das macht Mut.*

Analoges gilt in Beispiel 85 für den Adressierungsausdruck „@werkstatt_bpb" und für das Hashtag „#slml13", mit welchem vom Autor des Tweets eine thematische Vernetzung mit den Tweets anderer Nutzer erzeugt wird, die sich ebenfalls auf das Thema „slml13" beziehen.[4]

Vermutlich dürfte es sinnvoll sein, Einheiten dieser Art auf einer der Wortarten-annotation vor- oder nachgeordneten Verarbeitungs- und Modellierungsebene als Einheiten zu beschreiben, die zwischen der rein sprachlichen Ebene der Kommunikation und der hypermedialen Struktur der betreffenden Kommunikationsplattformen vermitteln und in denen die strategische Nutzung von Möglichkeiten zur technischen Vernetzung mit Formen der sprachlichen Referenzierung (von Themen und Adressaten) konvergieren.[5]

4 Automatische Verarbeitung sprachlicher Besonderheiten in der internetbasierten Kommunikation: Datengestützter Problemaufriss

In diesem Abschnitt geben wir einen Aufriss typischer Probleme bei der automatischen Wortartenannotation von Sprachdaten aus Genres internetbasierter Kommunikation mit Verarbeitungswerkzeugen, die von linguistischen Anwendern „off the shelf" über die Oberfläche der webbasierten Analyseplattform *WebLicht* (https://weblicht.sfs.uni-tuebingen.de/) genutzt werden können.

Wir werden zeigen, dass nur ein kleiner Teil der Herausforderungen, die mit der automatischen linguistischen Analyse solcher Daten verbunden sind, mit dem Fehlen geeigneter Kategorien im verwendeten POS-Tagset zu tun hat. Ein großer Teil der sprachlichen Einheiten ist auf Basis vorhandener POS-Kategorien – zumindest theoretisch – schon jetzt zutreffend klassifizierbar. Aufgrund nicht-standardkonformer Formmerkmale werden aber in vielen Fällen Tokens, die bei einer intellektuellen Klassifikation problemlos einer gängigen POS-Kategorie zugeordnet werden könnten, nicht zuverlässig erkannt. Weitere Verarbeitungsprobleme ergeben sich auf der Ebene der automatischen Tokenisierung und resultieren in Segmentierungen, deren Resultate sich auf höheren Verarbeitungsebenen nicht sinnvoll linguistischen Kategorien zuordnen lassen.

Zunächst beschreiben wir das für unsere Verarbeitungsexperimente zusammengestellte Evaluationsdatenset und die verwendeten Verarbeitungswerkzeuge. Anschließend fassen wir die im Zuge unserer Experimente festgestellten Verarbeitungsprobleme zu Problemtypen

zusammen. Dabei wird deutlich, auf welchen Ebenen des Verarbeitungsprozesses Bemühungen zur Verbesserung der Verarbeitungsergebnisse ansetzen können (und sollten) und für welche der in Abschnitt 3 vorgestellten Phänomene eine Erweiterung des POS-Tagsets erforderlich erscheint. In Abschnitt 5 werden wir davon ausgehend dann konkrete Vorschläge für Modifikationen und Erweiterungen des „Stuttgart-Tübingen Tagset" (STTS) formulieren.

4.1 Evaluationsdatenset und verwendete Sprachverarbeitungswerkzeuge

Das Evaluationsdatenset umfasst manuell zusammengestellte Belegsammlungen für ausgewählte Typen sprachlicher Phänomene in der internetbasierten Kommunikation. Berücksichtigt sind verschiedene Phänomentypen aus den in Abschnitt 3 unterschiedenen Phänomenbereichen *Geschriebene Umgangssprache*, *IBK-typische Akronyme* und *IBK-spezifische interaktive Einheiten*. Die Belege für die untersuchten Phänomene sind zu gleichen Teilen dem Dortmunder Chat-Korpus (Beißwenger 2013) und Diskussionsseiten der deutschsprachigen Wikipedia entnommen. Für jedes der Genres *Chat* und *Wikipedia-Diskussion* enthält das Datenset Subsets mit jeweils 100 Belegen für jeden der untersuchten Phänomentypen. Insgesamt umfasst das Evaluationsdatenset somit 1.000 Belege für das Vorkommen der untersuchten Phänomene (Tabelle 1).

Phänomentyp		**Belege**		
		Chat	Wikipedia-Diskussionen	DWDS
III.1+ III.3	Geschriebene Umgangssprache I: Umgangssprachlich fundierte Wortschreibungen und umgangssprachliche Lexik	100	100	100
III.2	Geschriebene Umgangssprache II: Kontraktierte Formen	100	100	
V	IBK-typische Akronyme	100	100	
VI.1	IBK-spezifische interaktive Einheiten I: Emoticons	100	100	
VI.2	IBK-spezifische interaktive Einheiten II: Aktionswörter	100	100	
Belege gesamt (nur IBK)		**1.000**		

Tab. 1: Zusammensetzung des Evaluationsdatensets.

Für den ersten Phänomentyp wurden zusätzlich in gleichem Umfang standardsprachliche Pendants im Kernkorpus des 20. Jahrhunderts des Projekts „Digitales Wörterbuch der Deutschen Sprache" (DWDS, Geyken 2007) erhoben. Ein Beleg enthält im Kontext eines Nutzerbeitrags (Posting) mindestens eine Instanz des jeweils untersuchten Phänomentyps (Beispiel 86). Nur für diesen Ausdruck wird im Folgenden der Output der Sprachverarbeitungswerkzeuge untersucht und bewertet.

(86) Beleg für die Verwendung eines Aktionsworts (**räusper**) aus dem Dortmunder Chat-Korpus nach Verarbeitung mit dem TreeTagger:

```
*räusper*/ADJA  Hömma/NN  woher/PWAV  kommste/VVFIN  denn/ADV ?/$.
Ick/ADJD  bin/VAFIN  aus/APPR  Do-Stadt/NN ,/$,  net/ADJD
aus/APPR  Berlin./NE
```

Die Verarbeitung des Testdatensets wurde mit Werkzeugen durchgeführt, die in der webbasierten Annotationsumgebung *WebLicht* (https://weblicht.sfs.uni-tuebingen.de/, Hinrichs/Zastrow/Hinrichs 2010) zur Verfügung stehen. *WebLicht* wurde im Rahmen des D-SPIN-Projekts maßgeblich am Seminar für Sprachwissenschaft der Universität Tübingen entwickelt und wird derzeit im Rahmen des Projekts CLARIN-D weiter ausgebaut.[6] Die Umgebung ermöglicht einen einfachen, Webservice-basierten Zugriff auf eine Vielzahl gängiger Sprachverarbeitungswerkzeuge, die somit nicht mehr lokal installiert und konfiguriert werden müssen, sondern direkt online aufgerufen und auf Daten angewendet werden können.

Die Tokenisierung und das POS-Tagging der Evaluationsdaten erfolgte in zwei voneinander unabhängigen Durchgängen mithilfe des TreeTaggers (Schmid 1994) und des POS-Taggers aus dem OpenNLP-Projekt (Modell: MaxEnt, Trainingsdaten: TIGER-Korpus; http://opennlp.apache.org) sowie den jeweils zugehörigen Tokenisierern (Abb. 3). Die Ergebnisse der Tokenisierung wurden zunächst separat evaluiert, um solche Probleme im Verarbeitungsprozess zu dokumentieren, die bereits auf Tokenisierungsebene auftreten. Anschließend wurde der Output der Tokenisierer für alle Datensätze manuell überprüft und nachbearbeitet, um als Input für das POS-Tagging normalisierte Daten zur Verfügung stellen zu können.

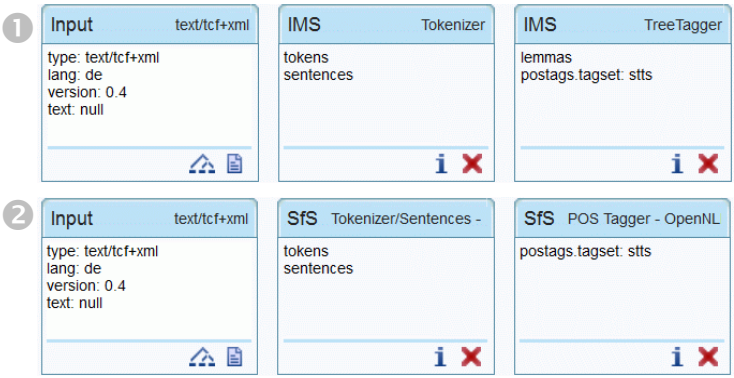

Abb. 3: Verwendete Verarbeitungsketten (Toolchains) und -werkzeuge, zusammengestellt über *WebLicht*.

4.2 Testergebnisse und Problemtypen

Bei den Tests der oben vorgestellten Verarbeitungsketten auf unseren Evaluationsdaten haben sich im Wesentlichen drei Einbruchstellen herauskristallisiert, an denen – teilweise in

Abhängigkeit zu bestimmten Phänomentypen – Verarbeitungsfehler entstehen und die somit für die Qualität der Tagging-Ergebnisse von zentraler Bedeutung sind. Ausgehend von diesen Einbruchstellen, die sich auf unterschiedliche Ebenen des Verarbeitungsprozesses und die darin genutzten Ressourcen beziehen, werden im Folgenden drei Problemtypen abgeleitet und auf der Basis der Testergebnisse konkretisiert.

4.2.1 Segmentierungsprobleme

Sprachliche Formen, die prinzipiell mithilfe vorhandener Kategorien im STTS klassifizierbar wären, werden teilweise schon bei der Wort- und Satzgrenzenerkennung als Einheiten repräsentiert, die sich nicht sinnvoll weiter analysieren lassen. Im einfachen Fall sind die Segmentierungsprobleme hauptsächlich durch die Auslassung von Spatien verursacht, die in Abschnitt 3 (Phänomentyp I.1) als typisches Schnellschreibphänomen beschrieben wurden. Da das Vorhandensein von Spatien für gängige Tokenisierer ein zentrales Kriterium für die Identifizierung von Tokengrenzen darstellt, kommt es in solchen Fällen entsprechend zu nicht sinnvollen Segmentierungen. Bei den Datensets zur geschriebenen Umgangssprache und zu IBK-typischen Akronymen, in denen 1–11% der Tokens vom Tokenisierer nicht korrekt zugeschnitten werden (s. Tabelle 2), stellen fehlende Spatien die Hauptursache für Segmentierungsfehler dar. Segmentierungsprobleme aufgrund fehlender Spatien illustriert Beispiel 87.

Eklatant schlechter werden die Segmentierungsergebnisse dann, wenn nicht nur die Tokengrenzen zum Problem werden, sondern wenn das Token als solches vom Tokenisierer nicht zuverlässig konstituiert werden kann. Im Falle von Emoticons wie z.B. „:-))“ oder „=o)“ werden bestimmte Typen von Zeichen – nämlich Interpunktions-, Buchstaben- und Sonderzeichen – zu Einheiten kombiniert, die nach den Segmentierungsregeln der Tokenisierer z.T. bereits selbst Tokenstatus haben können. Entsprechend führt die Anwendung dieser Regeln in vielen Fällen zu einer zu feinkörnigen Segmentierung, bei der Emoticons in Sequenzen von Interpunktions- und alphanumerischen Zeichen zerlegt werden (Beispiel 88). Im Chat-Subset wurden weniger als die Hälfte der untersuchten Formen dieses Typs korrekt segmentiert, im Subset mit den Wikipedia-Diskussionen sogar weniger als ein Viertel (s. Tab. 2).

Bei den Belegen mit Aktionswörtern, zu denen wir Inflektive und mehrteilige Inflektivkonstruktionen zählen, bereitete die Repräsentation der paarigen Asteriske, die diese Einheiten häufig umschließen (z.B. „*freu*“ oder „*baff bin*“), Probleme. Bei fast allen Belegen im Evaluationsdatenset werteten die Tokenisierer die Asteriske nicht als eigene Tokens, sondern als Bestandteile der adjazenten sprachlichen Ausdrücke (Beispiel 89).

(87) Segmentierungsprobleme: Der gesuchte Ausdruck „biste“ wurde aufgrund fehlenden Spatiums nicht als Token konstituiert, weiterer Fehler im Kontext (Output des Tokenisierers aus der TreeTagger-Toolchain):

```
wieso <token ID="t155">stoeps?biste</token> losgerannt einkaufen
udn ahst vergessen dich anzuziehen <token ID="t163">vorher?*G*
</token>
```

(88) Segmentierungsprobleme: Segmentierung eines Emoticons (Output des Tokenisierers aus der TreeTagger-Toolchain):

```
<token ID="t64">:</token>
<token ID="t65">)</token>
<token ID="t66">)</token>
<token ID="t67">)</token>
```

(89) Segmentierungsprobleme: Segmentierung eines mehrteiligen Aktionsworts (Output des Tokenisierers aus der TreeTagger-Toolchain):

```
<token ID="t946">*ins</token>
<token ID="t947">Bett</token>
<token ID="t948">fall*</token>
```

Phänomene	TreeTagger	OpenNLP-Tagger	Datenset
Geschriebene Umgangssprache I: Umgangssprachlich fundierte Wortschreibungen und umgangssprachliche Lexik	99 von 100	100 von 100	Wikipedia-Diskussionen
	91 von 100	92 von 100	Chat
	100 von 100	100 von 100	DWDS
Geschriebene Umgangssprache II: Kontraktierte Formen	100 von 100	100 von 100	Wikipedia-Diskussionen
	96 von 100	92 von 100	Chat
IBK-typische Akronyme	98 von 100	98 von 100	Wikipedia-Diskussionen
	89 von 100	92 von 100	Chat
IBK-spezifische interaktive Einheiten I: Emoticons	23 von 100	22 von 100	Wikipedia-Diskussionen
	48 von 100	45 von 100	Chat
IBK-spezifische interaktive Einheiten II: Aktionswörter	9 von 100	9 von 100	Wikipedia-Diskussionen
	0 von 100	0 von 100	Chat

Tab. 2: Korrekt segmentierte Instanzen von Emoticons und Aktionswörtern.

Die Segmentierungsprobleme in Bezug auf Emoticons und Aktionswörter waren insofern erwartbar, als die verwendeten Tokenisierer bislang nicht für den Umgang mit Phänomenen dieser Art angepasst wurden. Eine sinnvolle Tokenisierung solcher Einheiten ist keine triviale Aufgabe. Abhängig davon, ob eine Kombination aus Interpunktions- und alphanumerischen Zeichen in konventioneller Weise oder als Emoticon verwendet ist, muss entweder die Einheit als ganze oder jedes einzelne Zeichen als Token repräsentiert werden. Zudem existieren zu Emoticons homonyme Zeichenkombinationen, bei denen ein Teil der Sequenz den Status von Interpunktionszeichen, ein anderer Teil eine Funktion hat, die sich weder als Interpunktionszeichen noch als Emoticon fassen lässt. Beispiel 90 ist Storrer (2013) entnommen und entstammt der Wikipedia. Die Zeichenfolge repräsentiert ein noch ausstehendes Fußballergebnis; dass es sich weder um ein Emoticon noch um eine reine Sequenz von Interpunktionszeichen handelt, lässt sich nur durch eine Analyse des Kontexts disambiguieren.

(90) *Niederlande − Finnland* ***-:- (-:-)*** *Samstag, 29. August 2009, 17:30*
(Beispiel aus Storrer 2013)

4.2.2 Klassifizierungsprobleme

Von den *Segmentierungsproblemen*, die im Wesentlichen durch Schnellschreibphänomene und die für Tokenisierer „irreguläre" Nutzung von Interpunktions- und Sonderzeichen für den Aufbau IBK-spezifischer interaktiver Einheiten verursacht werden, unterscheiden wir die *Klassifizierungsprobleme*. Klassifizierungsprobleme ergeben sich auf der Ebene des POS-Tagging und bestehen darin, dass bestimmte, vom Tokenisierer korrekt als Tokens konstituierte Einheiten aufgrund nicht-standardkonformer Formmerkmale nicht mit dem Tag für die POS-Kategorie versehen werden können, der sie angehören.

Klassifizierungsprobleme ergeben sich im untersuchten Datenset zum einen für den Bereich der geschriebenen Umgangssprache (Phänomentyp III), zum anderen für die IBK-typischen Akronyme (Phänomentyp V).

Um die Klassifikation umgangssprachlich fundierter Wortschreibungen und umgangssprachlicher Lexik in unserem Evaluationsdatenset mit der Klassifikation entsprechender standardsprachlicher Formen zu vergleichen, haben wir ein Datenset mit standardsprachlichen Entsprechungen aus dem DWDS-Korpus zusammengestellt (s. Tabelle 1). Während die POS-Tagger die untersuchten Wortformen im DWDS-Datensatz in 87% (TreeTagger) bzw. 83% der Fälle (OpenNLP-Tagger) korrekt klassifizieren, erreichen sie für die entsprechenden Formen im IBK-Dastenset eine Genauigkeit von nur 34% und 44% (für die Belege aus Wikipedia-Diskussionen) bzw. 13% und 15% (für die Belege aus dem Dortmunder Chat-Korpus) (Tabelle 3). Beispiel 91 zeigt umgangssprachlich fundierte Wortschreibungen aus einer Wikipedia-Diskussionsseite und aus dem Dortmunder Chat-Korpus im Vergleich zu einer standardsprachlichen Entsprechung aus dem DWDS-Korpus plus die vom TreeTagger für die Wortformen jeweils vergebenen Tags:

(91) *Schaden kann **dat/ADJD** ja nich*
(Beispiel aus den Wikipedia-Diskussionen)

*syno **det/ADJA** is to wenig*
(Beispiel aus dem Dortmunder Chat-Korpus)

***das/PDS** ist schon kraftraubend*
(Beispiel aus dem DWDS-Korpus)

Im Falle der IBK-typischen Akronyme werden von den Taggern nur maximal 21% der untersuchten Formen korrekt zugeordnet (exemplarische Fehlanalysen s. Beispiel 92). Für Akronyme sind im STTS zwar keine eigenen Tags vorgesehen, die STTS-Guidelines sehen dafür aber eine Verfahrensweise vor:

> Abgekürzte Wortformen werden getaggt wie die ausgeschriebene Form. Mehrteilige, nicht durch Spatien getrennte Abkürzungen werden entsprechend ihrer syntaktischen Funktion klassifiziert. (Schiller et al. 1999: 9)

Um die Anwendung dieses Verfahrens auf IBK-typische Akronyme zu optimieren, müssen die Tagger auf den Umgang mit solchen Akronymen trainiert werden (z.B. auf Basis einer Liste).

(92) Klassifizierungsprobleme: POS-Tagging bei IBK-typischen Akronymen (Output des TreeTaggers):

***kA/NN** was das kostet*

***btw/ADJA** ... ward ihr denn auch alle fleißig wählen gewesen?*

***re/VVFIN** alle die ich eben schon begrüßt hatte.*

***Imho/NE** sind hier recht viele sogenannte Spaßformen aufgeführt*

Phänomentyp	TreeTagger	OpenNLP-Tagger	Datenset
Geschriebene Umgangssprache	34 von 100	44 von 100	Wikipedia-Diskussionen
	13 von 100	15 von 100	Chat
	87 von 100	83 von 100	DWDS
IBK-typische Akronyme	8 von 100	21 von 100	Wikipedia-Diskussionen
	10 von 100	17 von 100	Chat

Tab. 3: Korrekt vergebene POS-Tags für Fälle geschriebener Umgangssprache und für IBK-typische Akronyme.

4.2.3 Kategorienprobleme

Während im Falle von *Klassifizierungsproblemen* vorhandene Kategorien aufgrund von Formeigenschaften der analysierten Tokens nicht korrekt zugeordnet werden können, beruht die inkorrekte Klassifizierung von Tokens im Falle von *Kategorienproblemen* darauf, dass die relevanten Einheiten im Tagset überhaupt nicht kategorial repräsentiert sind. Entsprechend lassen sich Tokens, die diesen Einheiten zugehören, mit gängigen POS-Taggern also – erwartbar – noch gar nicht sinnvoll klassifizieren. Die Ursache des Problems liegt dabei auf der Ebene des Tagsets und nicht auf der Ebene der Tagger.

Kategorienprobleme bestehen im Falle der internetbasierten Kommunikation für zwei Typen von Einheiten: zum einen für die IBK-spezifischen interaktiven Einheiten (Emoticons, Aktionswörter, Adressierungen; Phänomentyp VI), zum anderen für die umgangssprachlichen kontraktierten Formen (Phänomentyp III.2). Letztere stellen zwar keinen IBK-spezifischen Phänomentyp dar, insofern sie auch in der gesprochenen Umgangssprache hoch frequent sind; in redigierten Texten, auf deren Verarbeitung die gegenwärtige Version des STTS primär optimiert ist, kommen sie aber bestenfalls in Ausnahmefällen vor.

Die jeweils 100 Instanzen von Emoticons und Aktionswörtern in unserem Evaluationsdatenset werden von den beiden Taggern mit ganz unterschiedlichen POS-Tags versehen:

Emoticons erhalten vom TreeTagger in Chats und Wikipedia-Diskussionen in den meisten Fällen Adjektiv- oder Nomen-Tags (*ADJA/D, NN/NE*). Der OpenNLP-Tagger hingegen behandelt die Einheiten häufig wie konventionelle Interpunktionszeichen – in den Belegen aus Wikipedia-Diskussionen sogar überwiegend, in den Chat-Belegen in etwa ähnlich häufig wie Nomen-Tags (Tabelle 4).

Aktionswörter, insbesondere einfache Inflektive, werden vom TreeTagger in 41% der Fälle als Verbformen klassifiziert. Existiert zu einem Aktionsausdruck eine homonyme Imperativform (z.B. **sing**, homonym zu *sing!*), wird diese in etwa der Hälfte der Fälle auch mit dem *VVIMP*-Tag versehen (Tabelle 5). Auch der OpenNLP-Tagger klassifiziert Aktionswörter häufig als Verbformen, vergibt allerdings am häufigsten (in 30% der Fälle) das Tag *XY* für „Nichtwort", das gemäß STTS-Guidelines „bei größeren Symbolgruppen, [...] sowie Kombinationen aus Ziffern und Zeichen, die sich nicht als *CARD* oder *ADJA* einordnen lassen", das Mittel der Wahl darstellt (Schiller et al. 1999: 74f.). Weiterhin vergeben beide Tagger für Aktionswörter häufig Adjektiv-Tags (Tabelle 4).

Phänomentyp	TreeTagger		OpenNLP-Tagger		Datenset
Emoticons	ADJA/D	52 von 100	$./$,/$(	68 von 100	Wikipedia-Diskussionen
	NN/NE	43 von 100	NN/NE	23 von 100	
	VVFIN	3 von 100	VV*	5 von 100	
	ADJA/D	52 von 100	NN/NE	40 von 100	Chat
	NN/NE	43 von 100	$./$(	36 von 100	
	CARD	4 von 100	XY	15 von 100	
Aktionswörter	VV*	41 von 100	XY	30 von 100	Wikipedia-Diskussionen
	NN/NE	32 von 100	VV*	24 von 100	
	ADJA/D	25 von 100	ADJA/D	22 von 100	
	VV*	41 von 100	XY	46 von 100	Chat
	ADJA/D	32 von 100	VV*	23 von 100	
	NN/NE	26 von 100	ADJA/D	20 von 100	

Tab. 4: Am häufigsten vergebene POS-Tags für Instanzen von Emoticons und Aktionswörtern.

Phänomentyp	TreeTagger	OpenNLP-Tagger	Datenset
Aktionswörter, zu denen es im verbalen Paradigma homonyme Imperativformen gibt	34 von 58	0 von 58	Wikipedia-Diskussionen
	28 von 47	1 von 47	Chat

Tab. 5: Vergebene VVIMP-Tags zu denjenigen Aktionswörtern, zu denen eine homonyme Imperativform existiert.

Auch für die Annotation umgangssprachlicher kontraktierter Formen gibt es im STTS derzeit noch keine geeignete Kategorie. Zwar existiert mit *APPRART* eine Kategorie für „Präpositionen mit inkorporiertem Artikel"; die Beschreibung der Kategorie (Schiller et al. 1999: 67) nennt als typische Vertreter aber Formen wie *am*, *zur*, *zum* und *ans*, die (auch) standardsprachlich etabliert sind, d.h. ihre umgangssprachliche Markiertheit verloren haben und überdies einen hohen Grammatikalisierungsgrad aufweisen. Zudem ist die Kategorie auf das Kontraktionsmuster Präposition + Artikel festgelegt. Umgangssprachlich werden Verschmelzungen aber auch nach diversen weiteren Mustern gebildet (vgl. die exemplarische Liste in Abschnitt 3, Phänomentyp III.2).

Das Subset „Kontraktierte Formen" aus unserem Evaluationsdatenset umfasst ausschließlich Formen mit verbaler erster Komponente. Am häufigsten werden diese von den beiden getesteten Taggern mit Tags für Verbformen versehen (*VVFIN*, *VAFIN*, *VMFIN*, *VVIMP*; Tabelle 6), wenngleich der TreeTagger fast ebenso häufig auch das NN-Tag vergibt. Eine künftige erweiterte Version des STTS sollte – wie dies auch von Gimpel et al. (2011) für das Englische unternommen wurde – eine eigene Kategorie für kontraktierte Formen vorsehen. Eine solche Kategorie könnte nicht nur für die linguistische Analyse von Sprachdaten aus Genres internetbasierter Kommunikation, sondern auch für das Tagging von Korpora mit Transkripten gesprochener Sprache von Nutzen sein.

Phänomentyp	TreeTagger		OpenNLP-Tagger		Datenset
Kontraktierte Formen	VVFIN	35 von 100	VVFIN	26 von 100	Wikipedia-Diskussionen
	VVIMP	1 von 100	VMFIN	10 von 100	
	NN	34 von 100	VAFIN	6 von 100	
	VVFIN	41 von 100	VVFIN	34 von 100	Chat
	VMFIN	6 von 100	VAFIN	11 von 100	
	NN	32 von 100	VMFIN	6 von 100	

Tab. 6: Am häufigsten vergebene POS-Tags für Instanzen von kontraktierten Formen.

5 Lösungsperspektiven und Vorschläge zur Modifikation des Stuttgart-Tübingen-Tagset (STTS) für die Annotation von Korpora internetbasierter Kommunikation

Die Ergebnisse aus den in Abschnitt 4 beschriebenen Tests zur automatischen Wortartenannotation von Sprachdaten aus Genres internetbasierter Kommunikation haben gezeigt, dass Probleme an unterschiedlichen Stellen des automatischen Verarbeitungsprozesses auftreten können:

1) **Segmentierungsprobleme** ergeben sich dadurch, dass auf der Ebene der automatischen Tokenisierung Zeichenfolgen als Tokens konstituiert werden, die in Verarbeitungsprozessen wie dem POS-Tagging, die tokenisierte Daten als Input nutzen, nicht sinnvoll weiter analysiert werden können. Wird das Tokenisierungsergebnis nicht zunächst manuell normalisiert, ergeben sich dadurch am Ende der Verarbeitungskette Fehler bei der Zuordnung von POS-Kategorien, die nicht dem POS-Tagger, sondern dem Tokenisierer anzulasten sind.

   Typische Ursachen für Segmentierungsprobleme sind die irreguläre Verwendung von Spatien sowie die Nutzung von Interpunktions- und Sonderzeichen für die Bildung von Emoticons und für die Kennzeichnung von Aktionswörtern.

   Lösungsmöglichkeiten für Probleme bei der automatischen Segmentierung sind entweder ein Training der Tokenisierungswerkzeuge auf den Umgang mit nicht-standardkonformer Schriftlichkeit und IBK-spezifischen lexikalischen Besonderheiten oder eine Normalisierung der Tokenisierungsergebnisse.

2) **Klassifizierungsprobleme** ergeben sich auf der Ebene des POS-Tagging und bestehen darin, dass bestimmte Tokens, für die im verwendeten POS-Tagsets geeignete Kategorien existieren, aufgrund nicht-standardkonformer Formmerkmale nicht mit dem entsprechenden Tag versehen werden können. Um Klassifizierungsprobleme als solche identifizieren zu können, muss sichergestellt sein, dass die Input-Daten für den Tagger eine Tokenisierung aufweisen, die frei von Segmentierungsproblemen ist.

   Ursachen für Klassifizierungsprobleme in Sprachdaten aus Genres internetbasierter Kommunikation sind u.a. Phänomene geschriebener Umgangssprache auf der Ebene der Orthographie und der Lexik sowie IBK-typische Akronyme. Fehler beim POS-Tagging, die sich durch Schnellschreibphänomene, Phänomene der graphischen Nachbildung suprasegmentaler Elemente der gesprochenen Sprache oder Verfremdungsschreibungen ergeben, fallen ebenfalls unter diesen Problemtyp.

Lösungsmöglichkeiten für Probleme bei der automatischen POS-Klassifizierung sind entweder ein Training der POS-Tagger auf den Umgang mit den entsprechenden Phänomenen oder eine Normalisierung der Input-Daten in einer dem POS-Tagging vorgeordneten Aufbereitungsphase.

3) Im Fall von **Kategorienproblemen** beruht die inkorrekte Zuordnung von POS-Kategorien zu Tokens darauf, dass für die Zielkategorien im Tagset keine Tags vorgesehen sind. Einheiten in Sprachdaten internetbasierter Kommunikation, für die im STTS bislang keine geeigneten Kategorien existieren, sind die IBK-spezifischen interaktiven Einheiten (mit den Subtypen Emoticons, Aktionswörter und Adressierungen) sowie umgangssprachliche kontraktierte Formen (*haste*, *biste*, *willste*, *machstes*; *aufm*; *isn* usw.).

   Durch Kategorienprobleme verursachte „Fehler" bei der Zuordnung von POS-Tags lassen sich nur durch eine Erweiterung des Tagsets lösen. Im Falle der kontraktierten Formen ist alternativ auch eine Normalisierung der Daten in einem Vorverarbeitungsprozess denkbar, bei dem die kontraktierten Formen entsprechend ihren standardsprachlichen Pendants (künstlich) aufgelöst werden (z.B. *biste* ⇨ *bist du*, *machstes* ⇨ *machst du es*, *aufm* ⇨ *auf dem*, *isn* ⇨ *ist denn* usw.) – dies aber verbunden mit einem Verlust der Information, dass es sich bei den dann künstlich erzeugten Varianten in den Originaldaten um typisch sprechsprachliche kontraktierte Formen handelt.

Die drei Problemtypen können einander überlagern. Abb. 4 ordnet die in Abschnitt 3 unterschiedenen Typen von sprachlichen Besonderheiten in der internetbasierten Kommunikation den verschiedenen Lösungsperspektiven zu.

Phänomentyp	**Bearbeitung der durch die Phänomene verursachten Verarbeitungsprobleme durch …**			
	Anpassung des Tokenisierers	Anpassung des POS-Taggers	Normalisierung der Tokenisierung	Erweiterung des Tagsets
I. Schnellschreibphänomene				
I.1 Irreguläre Verwendung von Spatien	√		√	
I.2 Tippfehler	√		√	
I.3 Ökonomiebedingte Abweichungen von den Normen für die Groß- und Kleinschreibung		√	√	
II. Graphische Nachbildung suprasegmentaler Elemente der gesprochenen Sprache		√	√	
III. Geschriebene Umgangssprache				
III.1 Umgangssprachlich fundierte Wortschreibungen		√	√	
III.2 Umgangssprachliche kontraktierte Formen			√	√
III.3 Umgangssprachliche Lexik		√		
IV. Verfremdungsschreibungen		√	√	
V. IBK-typische Akronyme		√		
VI. IBK-spezifische interaktive Einheiten				
VII.1 Emoticons	√	√		√
VII.2 Aktionswörter	√	√		√
VII.3 Adressierungen	√	√		√

Abb. 4: Sprachliche Besonderheiten in der internetbasierten Kommunikation und Perspektiven ihrer Behandlung bei der Verarbeitung mit Werkzeugen für die automatische linguistische Analyse.

Im Folgenden formulieren wir Vorschläge zur Modifikation und Erweiterung des STTS in Hinblick auf die Behandlung solcher Einheiten in der internetbasierten Kommunikation, die zum gegenwärtigen Stand der Kunst beim POS-Tagging Kategorienprobleme verursachen.

5.1 Emoticons, Aktionswörter, Adressierungen

Emoticons, Aktionswörter und Adressierungen können als durch die internetbasierte Kommunikation hervorgebrachte (Emoticons, Adressierungen) bzw. für eine breite Nutzung in schriftlicher interpersonaler Kommunikation adaptierte (Aktionswörter) Erweiterungen des lexikalischen Inventars gelten. Um sie in einem Kategoriensystem für das Tagging von Wortarten darstellbar zu machen, sind sie sinnvollerweise nicht als eine in der Luft hängende neue Kategorie zu konstituieren, sondern zu vorhandenen Kategorien in Beziehung zu setzen. Die Einordnung in einen grammatischen Beschreibungsrahmen sollte dabei den Funktionen Rechnung tragen, die sie in dialogischer Kommunikation übernehmen und mit denen sie die Möglichkeiten schriftlicher Kommunikation spezifisch erweitern.

In Abschnitt 3 (Phänomentyp VI) haben wir – ausgehend von den Vorschlägen in Beißwenger et al. (2012) – Emoticons, Aktionswörter und Adressierungen als IBK-spezifische Erweiterungen der Kategorie der *interaktiven Einheiten* dargestellt, die in der Konzeption der GDS (Zifonun et al. 1997) die Interjektionen (*ach*, *äh*, *mhm*, *ne*, *tja*) und die Responsive (*ja*, *nein*, *okay*) umfasst. Emoticons, Aktionswörter und Adressierungen erweitern diese Kategorie um Einheiten, die auf die Erfordernisse der Handlungskoordination und der emotionalen Kommentierung in dialogischer schriftlicher Kommunikation spezialisiert sind und die die Möglichkeiten schriftlicher Kommunikation im (zeitlichen und sozialen) Nahbereich ausbauen.

Um diese Einheiten in einer erweiterten Version des STTS-Kategoriensystems darstellbar zu machen, schlagen wir eine Restrukturierung desjenigen Systemausschnitts vor, der bislang durch die Kategorien *ITJ* (Interjektion) und *PTKANT* (Antwortpartikel) repräsentiert wird:

1. Die Kategorie *ITJ* verliert ihren Status als Hauptkategorie, die Hauptkategorie *PTK* (Partikeln) bleibt erhalten, wird aber um die Antwortpartikeln reduziert.
2. Auf Ebene der Hauptwortarten wird in Anlehnung an die entsprechende Kategorie der GDS das Konzept der *interaktiven Einheiten* als neue Hauptkategorie eingeführt. Die Kategorie wird durch ein eigenes Hauptkategorien-Tag dargestellt – zum Beispiel *IE*.
3. *ITJ* wird – bei gleichbleibender intensionaler und extensionaler Bestimmung – eine Subkategorie von *IE*. Um die kategoriale Einordnung im Tag selbst anzuzeigen, wird das Tag *ITJ* – entsprechend den Tagnamenkonventionen des STTS – zu *IEITJ* erweitert (lies: „Haupttyp: Interaktive Einheit, Subtyp: Interjektion“).
4. *PTKANT* wird – bei gleichbleibender intensionaler und extensionaler Bestimmung – ebenfalls eine Subkategorie von *IE*. Um die Kategorie der interaktiven Einheiten im STTS konsistent zur entsprechenden Kategorie der GDS darzustellen, wird der Terminus „Antwortpartikel“ ersetzt durch den Terminus „Responsiv“ (*RSP*). Um zudem

die kategoriale Einordnung im Tag selbst anzuzeigen, wird dafür das Tag *IERSP* vergeben (lies: „Haupttyp: Interaktive Einheit, Subtyp: Responsiv“).

Mit den Modifikationen 1–4 ist die Kategorie der interaktiven Einheiten im STTS etabliert. Die dabei vorgenommenen Restrukturierungen des Kategoriensystems sind aus unserer Sicht insofern moderat, als die schon existierenden Kategorien „Interjektion“ und „Antwortpartikel“ nur umbenannt, in ihrem Zuschnitt aber nicht verändert werden. Das STTS wird durch diese Restrukturierung vorbereitet, in einem weiteren Schritt die Emoticons, die Aktionswörter und die Adressierungen als IBK-spezifische Erweiterungen der interaktiven Einheiten aufzunehmen:

5. Einführung der neuen Kategorien *IEEMO* (Emoticons), *IEATW* (Aktionswort) und *IEADR* (Adressierung) als Subkategorien zu *IE* und auf gleicher Ebene wie *IEITJ* und *IERSP*.

Die entsprechenden Ausschnitte aus dem STTS-Kategoriensystem vor und nach der vorgeschlagenen Restrukturierung sind in Abb. 5 gegenübergestellt.

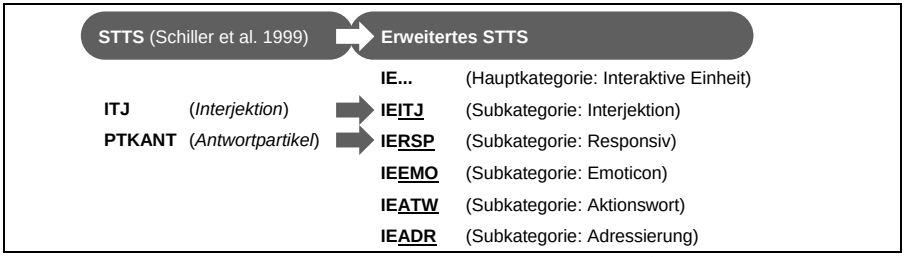

Abb. 5: Restrukturierung des STTS zur Darstellung IBK-spezifischer interaktiver Einheiten.

5.2 Umgangssprachliche kontraktierte Formen

Für die Darstellung umgangssprachlicher kontraktierter Formen (vgl. Phänomentyp III.2 in Abschnitt 3) bieten sich zwei verschiedene Lösungsvarianten an: eine einfache, bei welcher für Formen dieser Art eine neue Hauptkategorie (z.B. *KTR*) ohne weitere Subdifferenzierungen eingeführt wird, die kontraktive Formen aller möglichen Bildungsmuster umfasst. Diese einfache Lösung hätte den Vorteil, dass mögliche künftige Muster der Bildung von Verschmelzungen, die gegenwärtig nicht oder nur schwach produktiv sind, keiner erneuten Änderung des Tagsets bedürften, um im STTS dargestellt werden zu können. Der Nachteil dieser Lösung liegt entsprechend darin, dass das verwendete Tag wenig grammatische Strukturinformation enthält.

Die komplexere Lösungsvariante orientiert sich an der schon vorhandenen STTS-Kategorie *APPRART*, die (stark grammatikalisierte und auch standardsprachlich etablierte) Verschmelzungen des Typs Präposition+Artikel beschreibt und bei der sich das Bildungsmuster (Präposition+Artikel) aus der Benennung des Tags ableiten lässt. Nicht ablesen lässt sich aber, dass es sich bei den damit beschriebenen Einheiten um Verschmelzungen handelt.

Die komplexere Lösungsvariante für die Darstellung umgangssprachlicher kontraktierter Formen im STTS strebt an, wie im Falle von *APPRART* Strukturinformation zur Verschmelzung im Tag selbst zu kodieren. Darüber hinaus soll zusätzlich der Verschmelzungscharakter der beschriebenen Einheiten im Tag angezeigt werden. Dafür sind die folgenden Erweiterungen des Tagsets erforderlich:

1. Einführung einer neuen Hauptkategorie für umgangssprachliche kontraktierte Formen. Die Kategorie wird durch ein eigenes Hauptkategorien-Tag dargestellt – zum Beispiel *KTR*.
2. Einführung einer Reihe von Subkategorien für jeden einzelnen Strukturtyp der unter *KTR* erfassten Einheiten. Die Tags für die Subkategorien sind zusammengesetzt aus dem Namen der Hauptkategorie (*KTR*) als erstem Segment und den Namen derjenigen STTS-Kategorien als weiteren Segmenten, aus deren Einheiten die jeweiligen kontraktierten Formen gebildet sind. Für die in Abschnitt 3 exemplarisch aufgeführten Bildungsmuster ergeben sich somit z.B. die folgenden Kategorien und Tags:

– Präposition + Artikel (*innem*, *aufm*, *ausn*):[7]	*KTRAPPRART*
– Adverb + Artikel (*nochn*):	*KTRADVART*
– Konjunktion + Personalpronomen (*fallste*, *obse*):	*KTRKOPPER*
– Auxiliarverb + Personalpronomen (*haste*, *biste*):	*KTRVAPPER*
– Vollverb + Personalpronomen (*machste*, *gehste*, *denkste*):	*KTRVVPPER*
– Vollverb + zwei Personalpronomina (*machstes*, *gibstes*):	*KTRVVPPERPPER*
– Kopulaverb + Personalpronomen (*warens*):	*KTRVAPPER*
– Modalverb + Personalpronomen (*kannste*, *willste*, *sollste*):	*KTRVMPPER*
– Auxiliarverb + Abtönungspartikel (*isn*):	*KTRVAPTK*

Die vergebenen Tags wären in diesem Fall ungleich informativer – aber auch deutlich komplexer – als im Falle der zuerst vorgestellten Variante.

6 Fazit und Ausblick

Ziel dieses Beitrags war es, aus linguistischer Sicht die zentralen konzeptuellen Grundlagen bereitzustellen, die für eine Optimierung von Verfahren des Wortartentaggings für Zwecke der Annotation von Korpora zu Genres internetbasierter Kommunikation benötigt werden. Hierzu wurde zum einen eine Typologie sprachlicher Besonderheiten präsentiert, hinsichtlich derer sich die schriftliche Sprachverwendung in der internetbasierten Kommunikation charakteristischerweise von der Schriftlichkeit redigierter Texte (z.B. Zeitungstexte) unterscheidet (Abschnitt 3). In einem zweiten Schritt wurde eine Typologie von Problemen skizziert, die sich beim Wortartentagging von Daten aus Chats und aus Wikipedia-Diskussionsseiten ergeben (Abschnitt 4). Die Probleme betreffen unterschiedliche Ebenen und Ressourcen des Verarbeitungsprozesses; eine Optimierung automatischer Verfahren muss sowohl auf der Ebene der Tokenisierung wie auch auf der Ebene des POS-Taggings ansetzen und bedarf darüber hinaus eines Tagsets, das Erweiterungen für sprachliche Einheiten umfasst, die als spezifisch für die schriftliche internetbasierte Kommunikation gelten können. Ein Vorschlag, wie solche Erweiterungen in

einer modifizierten Version des STTS umgesetzt werden könnten, wurde in Abschnitt 5 formuliert.

Ausgehend von den vorgestellten Phänomen- und Problembeschreibungen und von der vorgeschlagenen STTS-Erweiterung können in einem nächsten Schritt manuell annotierte Trainigsdatensets (Goldstandard) zu Sprachdaten aus Genres internetbasierter Kommunikation aufgebaut und an diesen Sets existierende Tokenisierungs- und POS-Tagging-Verfahren retrainiert oder neu entwickelt werden. Entsprechende Vorhaben werden derzeit in unterschiedlichen Projektzusammenhängen verfolgt – u.a.:

a) im Projekt „Deutsches Referenzkorpus zur internetbasierten Kommunikation“ (*DeRiK*, Beißwenger et al. 2013), in dem derzeit ein linguistisch annotiertes Korpus mit deutschen Sprachdaten aus den wichtigsten Genres internetbasierter Kommunikation aufgebaut wird, das als eine Zusatzkomponente zu den Korpora geschriebener Sprache im Projekt „Digitales Wörterbuch der deutschen Sprache“ (http://www.dwds.de) konzipiert ist (Kooperationsprojekt des Instituts für deutsche Sprache der TU Dortmund mit der DWDS-Arbeitsgruppe an der Berlin-Brandenburgischen Akademie der Wissenschaften);

b) im DFG-Netzwerk „Empirische Erforschung internetbasierter Kommunikation“ (*Empirikom*, http://www.empirikom.net), in dem seit Frühjahr 2013 eine Shared Task zur automatischen Tokenisierung und Wortartenannotation von Sprachdaten aus der deutschsprachigen internetbasierten Kommunikation vorbereitet wird (Koordination: Michael Beißwenger);

c) im BMBF-Verbundprojekt „Korpus-basierte linguistische Recherche und Analyse mit Hilfe von Data-Mining“ (*KobRA*, http://www.kobra.tu-dortmund.de), in dem in einer Kooperation von Germanistischer Linguistik und Künstlicher Intelligenzforschung Verfahren des Maschinellen Lernens u.a. für die Adaption von POS-Tagging-Verfahren an die besonderen Anforderungen von Genres schriftlicher internetbasierter Kommunikation eingesetzt werden sollen (Projektleitung: Angelika Storrer).

Anmerkung

Wir danken den anonymen Reviewerinnen und Reviewern für ihre hilfreichen Anregungen und konstruktiven Kommentare zu einer früheren Version dieses Artikels.

Literatur

ANDROUTSOPOULOS, J./ZIEGLER, E. (2003). „Sprachvariation und Internet: Regionalismen in einer Chat-Gemeinschaft.“ In: Androutsopoulos, J./Ziegler, E. (Hrsg.): ‚Standardfragen‘. Soziolinguistische Perspektiven auf Sprachgeschichte, Sprachkontakt und Sprachvariation. Frankfurt: Peter Lang, 251-279.

AVONTUUR, T./BALEMANS, I./ELSHOF, L./VAN NOORD, N./VAN ZAANEN, M. (2012). „Developing a part-of-speech tagger for Dutch tweets.“ In: Computational Linguistics in the Netherlands Journal 2, 34–51.

BEIßWENGER, M. (2000). Kommunikation in virtuellen Welten: Sprache, Text und Wirklichkeit. Stuttgart: ibidem.

BEIßWENGER, M. (2003). „Sprachhandlungskoordination im Chat." In: Zeitschrift für germanistische Linguistik 31 (2), 198-231.

BEIßWENGER, M. (2007). Sprachhandlungskoordination in der Chat-Kommunikation. Berlin/New York: de Gruyter (Linguistik – Impulse & Tendenzen 26).

BEIßWENGER, M. (2013). „Das Dortmunder Chat-Korpus." In: Zeitschrift für germanistische Linguistik 41, H. 1, 161-164. Erweiterte Fassung online unter http://www.linse.uni-due.de/tl_files/PDFs/Publikationen-Rezensionen/Chatkorpus_Beisswenger_2013.pdf

BEIßWENGER, M./ERMAKOVA, M./GEYKEN, A./LEMNITZER, L./STORRER, A. (2012). „A TEI Schema for the Representation of Computer-mediated Communication." In: Journal of the Text Encoding Initiative (jTEI), Issue 3. http://jtei.revues.org/476 (DOI: 10.4000/jtei.476).

BEIßWENGER, M./ERMAKOVA, M./GEYKEN, A./LEMNITZER, L./STORRER, A. (2013). „DeRiK: A German Reference Corpus of Computer-Mediated Communication." In: Literary and Linguistic Computing 2013 (doi: 10.1093/llc/fqt038).

BEIßWENGER, M./STORRER, A. (Hrsg.; 2005). Chat-Kommunikation in Beruf, Bildung und Medien: Konzepte – Werkzeuge – Anwendungsfelder. Stuttgart: ibidem.

BEIßWENGER, M./STORRER, A. (2005a). „Chat-Szenarien für Beruf, Bildung und Medien." In: Beißwenger, M./Storrer, A. (Hrsg.): Chat-Kommunikation in Beruf, Bildung und Medien: Konzepte – Werkzeuge – Anwendungsfelder. Stuttgart: ibidem, 9-25.

BEIßWENGER, M./STORRER, A. (2012). „Interaktionsorientiertes Schreiben und interaktive Lesespiele in der Chat-Kommunikation." In: Zeitschrift für Literaturwissenschaft und Linguistik 168, 92-124.

BIBER, D. et al. (1999). Longman Grammar of Spoken and Written English. Edinburgh: Pearson Education Limited.

BIBER, D./CONRAD, S./LEECH, G. (2002). Longman Student Grammar of Spoken and Written English. Edinburgh: Pearson Education Limited.

BICK, E. (2010). „Degrees of Orality in Speech-like Corpora: Comparative Annotation of Chat and E-mail Corpora." In: Otoguro, R./Ishikawa. K./Umemoto, H./Yoshimoto, K./Harada, Y. (Hrsg.): Proceedings of the 24th Pacific Asia Conference on Language (PACLIC24). Institute for Digital Enhancement of Cognitive Development, Waseda University, 721-729.

BITTNER, J. (2003). Digitalität, Sprache, Kommunikation. Eine Untersuchung zur Medialität von digitalen Kommunikationsformen und Textsorten und deren varietätenlinguistischer Modellierung. Berlin (Philologische Studien und Quellen 178).

BLAKE, B.J. (2008). All About Language. New York: Oxford University Press.

BRINKER, K. (2001). Linguistische Textanalyse. Eine Einführung in Grundbegriffe und Methoden. 5., durchges. u. erg. Aufl. Berlin: Erich Schmitt Verlag (Grundlagen der Germanistik 29).

CRYSTAL, D. (2001). Language and the Internet. Cambridge: Cambridge University Press.

DUDEN-4[5] = DUDEN (1995). Die Grammatik. 5. Aufl. Mannheim: Bibliographisches Institut.

DUDEN-4[7] = DUDEN (2005). Die Grammatik. 7. Aufl. Mannheim: Bibliographisches Institut.

DÜRSCHEID, C. (2004). „Netzsprache – ein neuer Mythos?" In: Beißwenger, M./Hopffmann, L./Storrer, A. (Hrsg.): Internetbasierte Kommunikation (Osnabrücker Beiträge zur Sprachtheorie 68), 141–157.

DÜRSCHEID, C. (2005). „Normabweichendes Schreiben als Mittel zum Zweck." In: Muttersprache 115, 40-53.

DÜRSCHEID, C. (2005a). „Medien, Kommunikationsformen, kommunikative Gattungen." In: Linguistik online 22 (1). WWW-Ressource: http://www.linguistik-online.de/22_05/duerscheid.pdf.

EHLICH, K. (1983). „Text und sprachliches Handeln. Die Entstehung von Texten aus dem Bedürfnis nach Überlieferung." In: Assmann, A. et al. (Hrsg.): Schrift und Gedächtnis. Archäologie der literarischen Kommunikation I. München: Fink, 24-43.

EHLICH, K. (1984). „Zum Textbegriff." In: Rothkegel, A./Sandig, B. (Hrsg.): Text – Textsorten – Semantik. Hamburg: Buske, 9-25.

GDS = ZIFONUN, G./HOFFMANN, L./STRECKER, B. (1997). Grammatik der deutschen Sprache. 3 Bde. Berlin: de Gruyter (Schriften des Instituts für deutsche Sprache 7.1-7.3).

GEYKEN, A. (2007). „The DWDS corpus: A reference corpus for the German language of the 20th century." In: Fellbaum, C. (Hrsg.): Collocations and Idioms. London: continuum, 23-40.

GIESBRECHT, E./EVERT, S. (2009). „Part-of-speech tagging – a solved task? An evaluation of POS taggers for the Web as corpus." In: Alegria, I./Leturia, I./Sharoff, S. (Hrsg.): Proceedings of the 5th Web as Corpus Workshop (WAC5), San Sebastian, Spain. WWW-Ressource: http://cogsci.uni-osnabrueck.de/~severt/PUB/GiesbrechtEvert2009_Tagging.pdf

GIMPEL, K./SCHNEIDER, N./O'CONNOR, B. et al. (2011). „Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments." In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (HLT '11): short papers - Volume 2, 42–47. WWW-Ressource: http://dl.acm.org/ft_gateway.cfm?id=2002747&ftid=994684&dwn=1&CFID=247642444&CFTOKEN=38951532

GREENBAUM, S. (1996). The Oxford English Grammar. New York: Oxford University Press.

HERRING, S. C. (1999). „Interactional Coherence in CMC." In: Journal of Computer-Mediated Communication 4.4. WWW-Ressource: http://jcmc.indiana.edu/vol4/issue4/herring.html

HERRING, S. C. (Hrsg.) (1996). Computer-Mediated Communication. Linguistic, Social and Cross-Cultural Perspectives. Amsterdam/Philadelphia: John Benjamins (Pragmatics and Beyond New Series 39).

HERRING, S. C. (Hrsg.) (2010/11). Computer-Mediated Conversation, Part I/II. Special Issue of Language@Internet. WWW-Ressource: http://www.languageatinternet.org/articles/2010

HINRICHS, M./ZASTROW, T./HINRICHS, E. (2010). „WebLicht: Web-based LRT Services in a Distributed eScience Infrastructure." In: Proceedings of the Seventh conference on International Language Resources and Evaluation, Valetta, Malta.

JARBOU, S.O./AL-SHARE, B. (2012). „The Effect of Dialect and Gender on the Representation of Consonants in Jordanian Chat." In: Language@Internet 9. Online: http://www.languageatinternet.org/articles/2012/Jarbou

KILIAN, J. (2001). „T@stentöne. Geschriebene Umgangssprache in computervermittelter Kommunikation. In Chat-Kommunikation." In: Beißwenger, M. (Hrsg.): Sprache, Interaktion, Sozialität & Identität in synchroner computervermittelter Kommunikation. Perspektiven auf ein interdisziplinäres Forschungsfeld, Stuttgart: ibidem, 55-78.

KESTEMONT, M./PEERSMAN, C./DE DECKER, B./DE PAUW, G./LUYCKX, K./MORANTE, R./VAASSEN, F./VAN DE LOO, J./DAELMANS, W. (2012). „The Netlog Corpus. A Resource for the Study of

Flemish Dutch Internet Language.“ In: Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). Paris, 1569-1572.

KING, B. W. (2009). „Building and Analysing Corpora of Computer-Mediated Communication.“ In: Baker, P. (Hrsg.): Contemporary corpus linguistics. London: Continuum, 301-320.

KOCH, P./OESTERREICHER, W. (1994). „Schriftlichkeit und Sprache.“ In: Günther, H./Ludwig, O. (Hrsg.): Schrift und Schriftlichkeit. Ein interdisziplinäres Handbuch internationaler Forschung. Band 1. Berlin/New York: de Gruyter (Handbücher zur Sprach- und Kommunikationswissenschaft 12.1), 587-604.

LUCKHARDT, K. (2009). Stilanalysen zur Chat-Kommunikation. Eine korpusgestützte Untersuchung am Beispiel eines medialen Chats. Diss., TU Dortmund. Digitale Ressource: http://hdl.handle.net/2003/26055.

MCARTHUR, T. (Hrsg.) (1998). Concise Oxford Companion to the English Language. Oxford: Oxford University Press.

OWOPUTI, O./O'CONNOR, B./DYER, C./GIMPEL, K./SCHNEIDER, N. (2012). Part-of-Speech Tagging for Twitter: Word Clusters and Other Advances. Technical report, Carnegie Mellon University (CMU-ML-12-107). WWW-Ressource: http://ttic.uchicago.edu/~kgimpel/papers/CMU-ML-12-107.pdf

REYNAERT, M./OOSTDIJK, N./DE CLERCQ, O./VAN DEN HEUVEL, H./DE JONG, F.M.G. (2010). „Balancing SoNaR: IPR versus Processing Issues in a 500-Million-Word Written Dutch Reference Corpus.“ In: Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10). Paris: European Language Resources Association (ELRA), 2693-2698.

RITTER, A./CLARK, S./ETZIONI, M./ETZIONI, O. (2001). „Named Entity Recognition in Tweets: An Experimental Study.“ In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '11), 1524–1534. WWW-Ressource: http://dl.acm.org/ft_gateway.cfm?id=2145595&ftid=1146158&dwn=1&CFID=247642444&CFTOKEN=38951532

SCHIFFRIN, D. (1986). Discourse markers. Cambridge: Cambridge University Press (Studies in interactional sociolinguistics 5).

SCHILLER, A./TEUFEL, S./STÖCKERT, CH./THIELEN, CH. (1999). „Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset).“ Technischer Bericht. Universität Stuttgart, Institut für Maschinelle Sprachverarbeitung. WWW-Ressource: http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf

SCHLOBINSKI, P. (2001). „*knuddel – zurueckknuddel – dich ganzdollknuddel*. Inflektive und Inflektivkonstruktionen im Deutschen.“ In: Zeitschrift für germanistische Linguistik 29, 192-218.

SCHMID, H. (1994). „Probabilistic Part-of-Speech Tagging Using Decision Trees.“ In: Proceedings of International Conference on New Methods in Language Processing, Manchester, UK. WWW-Ressource: http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger1.pdf

SCHÖNFELDT, J./GOLATO, A. (2003). „Repair in Chats: A Conversation Analytic Approach.“ In: Research on Language and Social Interaction 36 (3), 241-284.

SCHWITALLA, J. (2006). Gesprochenes Deutsch. Eine Einführung,. 3., neu bearb. Aufl. Berlin: Erich Schmidt Verlag (Grundlagen der Germanistik 33).

STORRER, A. (2000). „Schriftverkehr auf der Datenautobahn. Besonderheiten der schriftlichen Kommunikation im Internet.“ In: Voß, G.G./Holly, W./Boehnke, K. (Hrsg.): Neue Medien im

Alltag: Begriffsbestimmungen eines interdisziplinären Forschungsfeldes. Opladen: Leske + Budrich, 153-177.

STORRER, A. (2001). „Getippte Gespräche oder dialogische Texte? Zur kommunikationstheoretischen Einordnung der Chat-Kommunikation." In: Sprache im Alltag. Beiträge zu neuen Perspektiven in der Linguistik. Herbert Ernst Wiegand zum 65. Geburtstag gewidmet. Hrsg. v. Andrea Lehr, Matthias Kammerer, Klaus-Peter Konerding, Angelika Storrer, Caja Thimm und Werner Wolski. Berlin: de Gruyter, 439-465.

STORRER, A. (2007). „Chat-Kommunikation in Beruf und Weiterbildung." In: Der Deutschunterricht, 49-61.

STORRER, A. (2009). „Rhetorisch-stilistische Eigenschaften der Sprache des Internets." In: Fix, U./Gardt, A./Knape, J. (Hrsg.): Rhetorik und Stilistik – Rhetorics and Stilistics. Ein internationales Handbuch historischer und systematischer Forschung. Berlin/New York: de Gruyter (Handbücher zur Sprach- und Kommunikationswissenschaft 31/2), 2211-2226.

STORRER, A. (2012). „Neue Text- und Schreibformen im Internet: Das Beispiel Wikipedia". In: Köster, J./Feilke, H./Steinmetz, M. (Hrsg.): Textkompetenzen in der Sekundarstufe II. Freiburg: Fillibach, 277-304.

STORRER, A. (2013). „Sprachstil und Sprachvariation in sozialen Netzwerken". In: Frank-Job, B./Mehler, A./Sutter, T. (Hrsg.): Die Dynamik sozialer und sprachlicher Netzwerke. Konzepte, Methoden und empirische Untersuchungen an Beispielen des WWW. Wiesbaden: VS Verlag für Sozialwissenschaften, 329-364.

STORRER, A. (im Druck). „Sprachverfall durch internetbasierte Kommunikation? Linguistische Erklärungsansätze – empirische Befunde." In: Sprachverfall? Dynamik – Wandel – Variation. Jahrbuch des Instituts für Deutsche Sprache 2013.

---

1 http://wiki.itmc.tu-dortmund.de/cmc/

2 http://www.tei-c.org/Activities/SIG/CMC/

3 Zu den kommunikativen Funktionen prosodischer Mittel in der gesprochenen Sprache vgl. z.B. Schwitalla (2006): 59-62.

4 „slml13" ist das Kürzel für eine Veranstaltung der Bundeszentrale für Politische Bildung mit dem Titel „SpeedLab: Mobiles Lernen – unabhängig von Raum und Zeit?", die am 26. April 2013 in Hannover stattfand.

5 Die Modellierung solcher „Linked-Data"-Phänomene steht u.a. auf der Agenda der Special Interest Group „Computer-Mediated Communication" im Rahmen der Text Encoding Initiative (http://www.tei-c.org/), die im Herbst 2013 ihre Arbeit aufgenommen hat.

6 Das Projekt D-SPIN („Deutsche Sprachressourcen-Infrastruktur") war bis 2011 der deutsche Beitrag zum europäischen CLARIN-Projekt („Common Language Resources and Technology Infrastructure"); seit 2011 werden die in D-SPIN begonnen Arbeiten im Nachfolgeprojekt CLARIN-D fortgeführt (vgl. http://www.d-spin.org/ und http://www.clarin-d.de/).

7 Erfasst würden unter dieser Kategorie dann lediglich die *umgangssprachlichen* Formen von Präposition-Artikel-Verschmelzungen (wie z.B. *aufm*, *innem*). Standardsprachlich etablierte Fälle des gleichen Bildungsmusters würden wie bisher unter *APPRART* erfasst.