Michael Beißwenger, Lothar Lemnitzer

Aufbau eines Referenzkorpus zur deutschsprachigen internetbasierten Kommunikation als Zusatzkomponente für die Korpora im Projekt „Digitales Wörterbuch der deutschen Sprache“ (DWDS)

Abstract

Dieser Beitrag gibt einen Überblick über die laufenden Arbeiten im Projekt „Deutsches Referenzkorpus zur internetbasierten Kommunikation“ (DeRiK), in dem ein Korpus zur Sprachverwendung in der deutschsprachigen internetbasierten Kommunikation aufgebaut wird. Das Korpus ist als eine Zusatzkomponente zu den Korpora im BBAW-Projekt „Digitales Wörterbuch der deutschen Sprache“ (*DWDS*, http://www.dwds.de) konzipiert, die die geschriebene deutsche Sprache seit 1900 dokumentieren.

Wir geben einen Überblick über die Motivation und Konzeption des Korpus sowie über die Projektziele (Abschnitte 2 und 3) und berichten über ausgewählte Anforderungen und Vorarbeiten im Zusammenhang mit der Korpuserstellung: a) die Integration des Korpus in die Korpusinfrastruktur des DWDS-Projekts (Abschnitt 4); b) die Entwicklung eines Schemas für die Repräsentation der strukturellen und linguistischen Besonderheiten von IBK-Korpora auf der Basis der Repräsentationsformate der *Text Encoding Initiative* (TEI-P5) (Abschnitt 5). Der Artikel schließt mit einer Skizze der Anwendungsszenarien für das Korpus in der korpusgestützten Sprachanalyse und der gegenwartssprachlichen Lexikographie (Abschnitt 6) sowie mit einem Ausblick (Abschnitt 7).

1. Einleitung

Dieser Beitrag gibt einen Überblick über die laufenden Arbeiten im Projekt „Deutsches Referenzkorpus zur internetbasierten Kommunikation“ (DeRiK), in dem in einer Kooperation der TU Dortmund und der Berlin-Brandenburgischen Akademie der Wissenschaften, Zentrum Sprache, seit 2010 ein Korpus zur Sprachverwendung in der deutschsprachigen internetbasierten Kommunikation aufgebaut wird. Am Projekt beteiligt sind neben den Verfassern dieses Beitrags Angelika Storrer (Dortmund) sowie Alexander Geyken und Maria Ermakova (Berlin). Die Basis des Korpus bilden nicht beliebige Webtexte, sondern die sprachlichen Äußerungen in solchen Webgenres, die in der englischsprachigen Forschung im Forschungsfeld „Computer-Mediated Communication“ (*CMC*) und in der deutschsprachigen Forschung unter dem Oberbegriff „Internetbasierte Kommunikation“ (*IBK*) untersucht werden.[1] Im Fokus stehen dabei Kommunikationstechnologien, die auf der Infrastruktur des Internet und seiner Dienste aufsetzen und die für die Realisierung

1 Überblicke zu den sprachlichen und kommunikativen Besonderheiten internetbasierter Kommunikation bieten z.B. Herring (1996; 2010/2011), Crystal (2001; 2011), Runkehl et al. (1998), Beißwenger (2001), Beißwenger & Storrer (2008) und Storrer (2013).

dialogischer interpersonaler Kommunikation konzipiert sind. Prominente Beispiele für Genres internetbasierter Kommunikation sind Chats und Instant-Messaging-Dialoge, Diskussions-Threads in Online-Foren und in Wikis, Threads mit Nutzerkommentaren in Weblogs, Videoplattformen (z.B. *YouTube*) und auf den Profilseiten sozialer Netzwerke (z.B. *Facebook*), die Kommunikation anhand von *Twitter*-Postings (Tweets) sowie in multimodalen Kommunikationsumgebungen wie *Skype*, MMORPGs („Massively Multi-Player Online Role Playing Games") und in „virtuellen Welten" (*SecondLife* u.a.).

Der Fokus von DeRiK liegt auf der schriftlichen Sprachverwendung in der internetbasierten Kommunikation. Das Korpus ist als eine Zusatzkomponente zu den Korpora im BBAW-Projekt „Digitales Wörterbuch der deutschen Sprache" (*DWDS*, http://www.dwds.de) konzipiert, die die geschriebene deutsche Sprache seit 1900 dokumentieren.

In den folgenden Abschnitten geben wir einen Überblick über die Motivation und Konzeption des Korpus sowie über die Projektziele (Abschnitte 2 und 3) und berichten über ausgewählte Anforderungen und Vorarbeiten im Zusammenhang mit der Korpuserstellung:

- die Integration des Korpus in die Korpusinfrastruktur des DWDS-Projekts (Abschnitt 4);
- die Entwicklung eines Schemas für die Repräsentation der strukturellen und linguistischen Besonderheiten von IBK-Korpora auf der Basis der Repräsentationsformate der *Text Encoding Initiative* (TEI-P5) (Abschnitt 5).

Der Artikel schließt mit einer Skizze der Anwendungsszenarien für das Korpus in der korpusgestützten Sprachanalyse und der gegenwartssprachlichen Lexikographie (Abschnitt 6) sowie mit einem Ausblick (Abschnitt 7).

2. Motivation: Ein Blick in die Korpuslandschaft

Gegenwärtig gibt es erst wenige Korpora zu Genres internetbasierter Kommunikation ('IBK-Korpora'). Bei den meisten existierenden Korpora mit IBK-Daten handelt es sich um Ressourcen, die für die interne Nutzung in einzelnen Forschungsprojekten aufgebaut wurden und die für die Fachcommunitys nicht ohne Weiteres zugänglich sind. Die unbefriedigende Abdeckung des wichtigen Kommunikationsbereichs „internetbasierte Kommunikation" in der Korpuslandschaft zur deutschen Gegenwartssprache (wie auch zu vielen anderen Sprachen) ist darauf zurückzuführen, dass in Bezug auf die Erhebung, Dokumentation, linguistische Annotation und Beschreibung von IBK-Daten wie auch auf Aspekte ihrer Bereitstellung für Forschungs- und Lehrzwecke derzeit noch viele offene Fragen und Herausforderungen bestehen. Forschungs- und Klärungsbedarf besteht u.a. hinsichtlich der folgenden Punkte:

- **Rechtliche und ethische Fragen:** Unter welchen Bedingungen können Sprachdaten aus Genres internetbasierter Kommunikation für die wissenschaftliche Nutzung archiviert, für linguistische Analysezwecke aufbereitet und annotiert und im Rahmen von Korpora als Forschungsressource bereitgestellt werden?
- **Fragen der Strukturrepräsentation und der Interoperabilität:** Welche Formate für die Repräsentation von Textgenres lassen sich sinnvoll für die Strukturbe-

schreibung von IBK-Korpora adaptieren? Wie müssen existierende Repräsentationsschemata angepasst werden, um die Struktur von Threads und Logfiles darzustellen? Welche Modellierungseinheiten können aus existierenden Schemata übernommen werden und für welche Einheiten bedarf es neuer, IBK-spezifisch angepasster Modelle?

- **Fragen der linguistischen Annotation:** Wie können Werkzeuge für die automatische Sprachverarbeitung (für die Tokenisierung, Normalisierung, Lemmatisierung, das Part-of-speech-Tagging und die syntaktische Annotation) sowie die von ihnen verwendeten Tagsets für den Umgang mit orthographischen Normabweichungen und IBK-spezifischen Stilelementen (z.B. Emoticons, Aktionswörter, Hashtags, Adressierungen) sowie für die Behandlung von Phänomenen der konzeptionellen Mündlichkeit angepasst werden?

- **Fragen der Integration von IBK-Ressourcen in bestehende Korpusinfrastrukturen:** Wie können IBK-Daten in existierende Infrastrukturen für die Verwaltung, Bereitstellung und Abfrage von Sprachkorpora integriert werden? Wie lassen sie sich vergleichend mit anderen Typen von Korpora (Textkorpora, Korpora gesprochener Sprache) nutzen und analysieren? Wie sind IBK-Daten an der Nutzerschnittstelle von Korpusabfragesystemen zu präsentieren? Wie können IBK-Korpora anhand von Metadaten beschrieben werden?

Diese und weitere Herausforderungen beim Aufbau von IBK-Korpora sind in der Forschung bekannt (vgl. z.B. Beißwenger & Storrer 2008, King 2009, Storrer 2013) und aktuell Thema verschiedener Netzwerke und Arbeitsgruppen – u.a. des DFG-Netzwerks „Empirische Erforschung internetbasierter Kommunikation“ (*Empirikom*, http://www.empirikom.net) und des GSCL-Arbeitskreises *Social Media /Internetbasierte Kommunikation* (http://gscl.org/ak-ibk.html) –, für ihre Bearbeitung gibt es bislang aber erst wenige „Best Practice“-Beispiele. Einen Überblick über IBK-Korpora auf dem Stand von 2008 geben Beißwenger & Storrer (2008). Beispiele für das Englische sind das *NPS Chat Corpus* (Forsyth et al. 2007) sowie das *Queer Chat-Room Corpus* (King 2009), für das Flämische das *Netlog Corpus* (Kestemont et al. 2012). Für das Deutsche existiert mit dem von 2002 bis 2008 an der TU Dortmund aufgebauten *Dortmunder Chat-Korpus* eine Ressource zur Sprachverwendung und sprachlichen Variation in der Chat-Kommunikation, die seit 2005 auch online zur Verfügung steht (http://www.chatkorpus.tu-dortmund.de, Beißwenger 2013). Für das Französische gibt es mit den *Learning and Teaching Corpora* (*LETEC*) eine Sammlung mit Daten aus multimodalen Lehr-/Lernarrangements (Reffay et al. 2012); mit *CoMeRe* (*Corpus de communication médiée par les réseaux*, http://comere.org) befindet sich darüber hinaus gegenwärtig ein genreheterogenes Korpus zur französischsprachigen internetbasierten Kommunikation im Aufbau. Im Projekt *Web2Corpus_it* (*Corpus Italiano di Communicazione mediata dal Computer,* http://www.glottoweb.org/web2corpus/) soll ein ausgewogenes IBK-Korpus für das Italienische entstehen. Beispiele für Referenzkorpora zur Gegenwartssprache, die Teilkorpora zu IBK-Genres integrieren, sind das niederländische *SoNaR*-Korpus (*Stevin Nederlandstalig Referentiecorpus*, Reynaert et al. 2010, Oostdijk et al. 2013)

sowie für das Estnische das *National Corpus of Estonian Language* (http://www.cl.ut.ee/korpused/segakorpus/).

Ein wichtiges Desiderat für das Deutsche ist ein Referenzkorpus zur internetbasierten Kommunikation, das für linguistische Analysezwecke aufbereitet ist, in einem anerkannten Repräsentationsformat zur Verfügung gestellt wird, vergleichende Analysen mit der Sprachverwendung in redigierten Texten (Textkorpora) ermöglicht und das in Forschung und Lehre als Basis für empirische Untersuchungen zur sprachlichen Variation in der internetbasierten Kommunikation sowie für die datengestützte Vermittlung ihrer sprachlichen und kommunikativen Besonderheiten genutzt werden kann. Das DeRiK-Projekt strebt an, einen Beitrag zur Schließung dieser Lücke in der Korpuslandschaft zur deutschen Gegenwartssprache zu leisten.

3. Zur Konzeption des Korpus

Ziel des DeRiK-Projekts ist der Aufbau einer Zusatzkomponente zu den DWDS-Korpora, die die deutschsprachige internetbasierte Kommunikation dokumentiert und durch deren Integration in die DWDS-Korpusinfrastruktur es u.a. möglich werden soll, die schriftliche Sprachverwendung im Internet vergleichend mit der schriftlichen Sprachverwendung in redigierten Texten zu untersuchen. Redigierte Texte sind in den Korpora des DWDS-Projekts bereits umfangreich dokumentiert: Das Kernkorpus umfasst Texte aus den Textsortenbereichen Belletristik, Wissenschaft, Zeitung und Gebrauchstexte für alle Dekaden seit 1900 (vgl. Geyken 2007). Die schriftliche Sprachverwendung im Netz ist in den Korpora hingegen bislang nicht berücksichtigt. Die DeRiK-Komponente soll diesen wichtigen Kommunikationsbereich nicht nur für die empirische Forschung zu IBK-Phänomenen, sondern auch für den Bereich der lexikographischen Bearbeitung der deutschen Gegenwartssprache korpuslinguistisch erschließen.

Unter der letztgenannten Perspektive erweitert DeRiK die Ressourcen für die Erarbeitung des DWDS-Wörterbuchs in zweierlei Weise: Zum einen wird das Korpus zahlreiche Beispiele für schriftliche Sprachverwendung im Duktus der konzeptionellen Mündlichkeit (i.S.v. Koch & Oesterreicher 1994) umfassen. Dies ermöglicht es, in die Wörterbuchartikel auch Belegbeispiele aus Kontexten informeller Schriftlichkeit aufzunehmen, was insbesondere für die Beschreibung umgangssprachlich markierter Lexik oder von sprachlichen Einheiten, die wichtige Aufgaben bei der Handlungskoordination in Gesprächen übernehmen (z.B. Interjektionen), bedeutsam ist. Zum anderen ist zu erwarten, dass in Daten zur Sprachverwendung in IBK-Genres zahlreiche Wörter und Wortbedeutungen dokumentiert sind, die mit traditionellen Quellen arbeitenden Lexikographen vermutlich entgehen (vgl. die Beispiele in Abschnitt 4). Ein regelmäßig aktualisiertes, hinsichtlich der Genres breit gestreutes Referenzkorpus zur internetbasierten Kommunikation stellt somit für die korpusbasierte Lexikographie eine wichtige Ergänzung zu den bisher typischerweise verwendeten Ressourcen dar. Durch ihre Verfügbarkeit kommt man dem Versprechen der gegenwartssprachlichen Lexikographie, den Wortschatz einer Sprache in seiner Gänze zu beschreiben, näher.

DeRiK ist konzipiert als

- ein **Referenzkorpus** zur internetbasierten Kommunikation, das der Fachcommunity als Basis für korpusgestützte Untersuchungen und für die Vermittlung der

sprachlichen Besonderheiten von IBK-Genres in der Lehre zur Verfügung gestellt wird;

- ein **ausgewogenes Korpus**, das – soweit rechtlich möglich – Daten aus den meistgenutzten IBK-Genres umfasst und die einzelnen Genres nach Popularität gegeneinander gewichtet;
- ein **zeitlich gestaffeltes Korpus**, bei dem – analog zur Datenerhebung für das DWDS-Kernkorpus – die Datenerhebung nicht nur einmalig, sondern mehrfach in regelmäßigen Abständen erfolgen soll, wodurch es möglich wird, auch sprachlichen Wandel *innerhalb* der internetbasierten Kommunikation darzustellen;
- ein **annotiertes Korpus**, das neben einer linguistischen Basisannotation auch Annotationen zu charakteristischen sprachlichen und strukturellen Besonderheiten bei der Sprachverwendung im Netz umfasst.

Die anvisierte Größe des Korpus sind 10 Millionen Token je Dekade, beginnend mit dem Jahr 2010 (dem Jahr des Projektbeginns). Für die Zusammensetzung des Korpus wurde ein Idealschlüssel entwickelt, der von den Ergebnissen der jährlich durchgeführten *ARD/ZDF-Onlinestudie* (http://www.ard-zdf-onlinestudie.de/, vgl. van Eimeren & Frees 2013) ausgeht und aus den in der Studie beschriebenen Präferenzen deutscher Internetnutzer für internetbasierte Kommunikationstechnologien und der Online-Affinität verschiedener Altersgruppen einen Faktor für die Gewichtung unterschiedlicher IBK-Genres bei der Festlegung der Datensets für einen Erhebungszeitraum ableitet. Der Schlüssel nutzt verschiedene Teilergebnisse der Studie als Grundlage: zum einen die Verbreitung der Internetnutzung nach Altersgruppen (vgl. Tab. 1), zum anderen die Präferenzen der Nutzer für bestimmte Typen von Online-Anwendungen (vgl. Tab. 2).

	1997	2000	2003	2006	2009	2010	2011	2012	2013
Gesamt	6,5	28,6	53,5	59,5	67,1	69,4	73,3	75,9	77,2
14-19 J.	6,3	48,5	92,1	97,3	97.5	100,0	100,0	100,0	100,0
20-29 J.	13,0	54,6	81,9	87,3	95,2	98,4	98,2	98,6	97,5
30-39 J.	12,4	41,1	73,1	80,6	89,4	89,9	94,4	97,6	95,5
40-49 J.	7,7	32,2	67,4	72,0	80,2	81,9	90,7	89,4	88,9
50-59 J.	3,0	22,1	48,8	60,0	67,4	68,9	69,1	76,8	82,7
ab 60 J.	0,2	4,4	13,3	20,3	27,1	28,2	34,5	39,2	42,9

Tab. 1: Internetnutzer in Deutschland 1997 bis 2013 nach Alter: mindestens gelegentliche Nutzung, in %. Quelle: http://www.ard-zdf-onlinestudie.de/index.php?id=421

	Gesamt	14-29	30-49	50-69	ab 70
Suchmaschinen nutzen	83	90	87	76	61
senden/empfangen von E-Mails	79	80	85	73	64
zielgerichtet bestimmte Angebote/informationen suchen	72	80	77	64	50
einfach so im Internet surfen	44	57	45	35	22
Onlinecommunitys nutzen	39	76	38	13	7
sog. "Apps" auf Mobilgeräten nutzen, um ins Internet zu gehen	35	60	35	17	8
Homebanking	34	33	39	31	31
Videoportale nutzen	32	65	28	11	7
Chatten	26	59	20	9	3
Herunterladen von Dateien	23	35	22	15	6
Kartenfunktionen nutzen	20	27	20	15	10
Onlinespiele	16	23	17	9	7
Audios im Internet herunterladen/anhören	14	31	12	5	0
Musikdateien aus dem Internet	14	33	9	4	0
Video/TV zeitversetzt	13	24	11	11	4
live im Internet Radio hören	13	22	11	8	2
RSS-feeds/Newsfeeds	10	18	10	4	4
Gesprächsforen	10	15	12	4	2
Ortungsdienste für ortsbezogene Informationen nutzen	10	14	8	9	5

Tab. 2: Onlineanwendungen 2013 nach Alter: mindestens einmal wöchentlich genutzt, in % (Ausschnitt). Quelle: http://www.ard-zdf-onlinestudie.de/index.php?id=423. Anwendungstypen mit Relevanz für DeRiK (= internetbasierte Kommunikationstechnologien oder Anwendungen mit entsprechenden Funktionen) sind in der Tabelle hervorgehoben.

Die Nutzung von Onlinecommunitys wurde in der ARD/ZDF-Onlinestudie zudem auch in einer Teilstudie zu den Präferenzen für ausgewählte „Social Media"- bzw. „Web 2.0"-Anwendungen abgefragt. In dieser Teilstudie wird weiter differenziert nach privaten und

beruflichen Communitys; daneben wird die Nutzung der Wikipedia, von Weblogs und von Twitter sowie von Videoportalen und Fotocommunitys dargestellt (vgl. Tab. 3 sowie im Detail Busemann 2013). Der ideale Schlüssel für DeRiK sieht vor, die in den Tab. 2 und 3 dargestellten Nutzungspräferenzen für die verschiedenen Altergruppen mit einem Faktor zu multiplizieren, der der Online-Affinitität der jeweiligen Altersgruppe entspricht und der sich aus den in Tab. 1 dargestellten Zahlen ableitet. So würde beispielsweise die Präferenz der 14-29-Jährigen für die Nutzung von Chats in der Studie aus 2013 (59%, s. Tab. 2) mit dem Faktor 0,9875 multipliziert (= Durchschnitt aus 100% und 97,5% Online-Nutzung bei der Altersgruppe der 14-19- und der 20-29-Jährigen, Tab. 1), die Präferenz der 30-49-Jährigen (20%) hingegen mit dem Faktor 0,922 (= Durchschnitt aus 95,5% und 88,9% Online-Nutzung bei der Altersgruppe der 30-39- und der 40-49-Jährigen) usf. Der exakte Anteil von Daten aus einem bestimmten Anwendungstyp (z.B. Chats) in der Gesamtdatenmenge für einen Erhebungszeitraum ergäbe sich schließlich aus dem Durchschnitt der gewichteten Präferenzen aller Altersgruppen im Verhältnis zu den ermittelten Werten für alle anderen relevanten Anwendungstypen. Für Anwendungstypen, deren Präferenzen in beiden Teilstudien der ARD/ZDF-Onlinestudie abgefragt wurden (Onlinecommunitys, Videoportale) sollen dabei jeweils die Werte aus der Teilstudie „Onlineanwendungen“ zugrunde gelegt werden.

In Bezug auf die in Tab. 3 dargestellten Werte zur Nutzung von Wikipedia, Weblogs und Twitter ist zu bedenken, dass bei allen drei (Typen von) Anwendungen die rezeptive Nutzung höher ist als die produktive Nutzung: Viele Onliner nutzen die Wikipedia als Nachschlagewerk, nur ein Teil der Nachschlagenden tritt aber selbst als Autor von Wikipedia-Artikeln und als Diskutant auf Diskussionsseiten in Erscheinung; dennoch können die Diskussionsseiten auch von nicht aktiv zum Ausbau der Anwendung beitragenden Nutzern eingesehen werden. Tweets, die in Twitter gepostet werden, können von einer Vielzahl von Nutzern (auch solchen, die nicht selbst aktiv „twittern“) gelesen werden; Gleiches gilt für Kommentare zu Weblog-Einträgen. Da die Frage der Gewichtung der rein rezeptiven gegenüber der auch produktiven Nutzung in Bezug auf Wikipedia, Twitter und Weblogs schwierig zu beantworten ist, ist bislang vorgesehen, die bei der Berechnung des Idealschlüssels unberücksichtigt zu lassen: Da in der Wikipedia geführte schriftliche Diskussionen sowie Tweets und Weblog-Kommentare auch für nur rezeptiv Zugreifende jederzeit einsehbar sind und ein Wechsel von der ausschließlich rezeptiven zur auch produktiven Nutzung der benannten Anwendungstypen jederzeit möglich ist, wird hier vorerst nicht weiter differenziert.

Bei der Entscheidung, aus welchen Online-Anwendungen konkret Daten für die einzelnen Anwendungstypen erhoben werden, soll Wert darauf gelegt werden, Vielfalt – und damit sprachliche Variation bei der Nutzung ein- und desselben Anwendungstyps – abzubilden. So wird beispielsweise für Chats angestrebt, die Daten nicht sämtlich aus demselben Chat-Angebot zu erheben und zudem die Nutzung von Chats in unterschiedlichen Handlungsbereichen abzubilden (Freizeitkommunikation, berufliche Nutzung, Nutzung in Lehr/Lernkontexten). Analog soll auch bei allen anderen Anwendungstypen verfahren werden.

Da die Onlinestudie jährlich aktualisiert wird, wird es möglich sein, den Schlüssel je Erhebungszeitraum an die jeweils aktuellen Zahlen anzupassen und dabei ggf. auch neu aufkommende IBK-Genres zu berücksichtigen.

	Gesamt	14-19	20-29	30-39	40-49	50-59	60-69	ab 70
Wikipedia	74	95	93	81	77	61	47	32
Videoportale (z.B. YouTube)	60	91	87	71	62	43	25	13
private Netzwerke u. Communitys	46	87	80	55	38	21	16	6
Fotosammlungen, Communitys	27	28	38	37	26	16	17	13
berufliche Netzwerke u. Communitys	10	5	14	19	13	4	2	0
Weblogs	16	18	31	19	17	7	3	5
Twitter	7	22	10	7	5	3	4	0

Tab. 3: Nutzung von Web 2.0-Anwendungen nach Alter 2013: gelegentliche Nutzung, in %. Quelle: http://www.ard-zdf-onlinestudie.de/index.php?id=397

Der Idealschlüssel wird bei der Datenerhebung allerdings nur mit Einschränkungen umgesetzt werden können: Aufgrund der unklaren Rechtslage in Bezug auf die Erhebung, Aufbereitung, Bereitstellung und Nutzung von IBK-Daten in Korpora werden bis auf Weiteres für das Korpus nur Daten aus solchen Kommunikationsumgebungen im Netz erhoben werden können, bei denen die Nutzung unproblematisch bzw. durch explizit deklarierte Lizenzen geregelt ist – u.a. Daten aus Anwendungen, deren Inhalte unter CC-BY-SA („Creative Commons: Namensnennung – Weitergabe unter gleichen Bedingungen“[2]) oder vergleichbaren Modellen für eine Bearbeitung und Weitergabe lizenziert sind.

4. Integration des Korpus als Zusatzkomponente in das DWDS-Korpusframework

Das Digitale Wörterbuch der deutschen Sprache ist ein digitales lexikalisches System, das an der Berlin-Brandenburgischen Akademie der Wissenschaften entwickelt wurde und weiterentwickelt wird. Das System eröffnet den Nutzern einen integrierten Zugang zu drei verschiedenen Typen von Ressourcen (Geyken 2007, Klein & Geyken 2010):

2 http://creativecommons.org/licenses/by-sa/2.0/de/legalcode (der Lizenztext) und http:// creativecommons.org/licenses/by-sa/2.0/de/ (eine verständliche Zusammenfassung des Inhalts).

a) **Lexikalische Ressourcen:** ein gegenwartssprachliches Wörterbuch, basierend auf dem retrodigitalisierten *Wörterbuch der deutschen Gegenwartssprache* (WDG, Klappenbach & Steinitz 1964-1977), das *Etymologische Wörterbuch des Deutschen* (Pfeifer 1993), die Erstbearbeitung des *Deutschen Wörterbuchs* von Jacob Grimm und Wilhelm Grimm (Jacob Grimm & Wilhelm Grimm, 1852-1971) sowie ein Thesaurus *(Openthesaurus)*.

b) **Korpora:** Das DWDS bietet ein ausgewogenes Korpus des 20. und des frühen 21. Jahrhunderts und darüber hinaus Zeitungskorpora und Spezialkorpora. Die jüngsten Texte stammen momentan aus dem Jahr 2010.

c) **Statistische Ressourcen für Wörter und Wortkombinationen:** Angeboten werden Wortprofile und Wortverlaufskurven für den gegenwartssprachlichen Wortschatz; die Auswertungen basieren auf dem Kernkorpus und den Zeitungskorpora.

Ergebnisse zu Suchanfragen auf diesen Ressourcen werden in einer panelbasierten, nutzerkonfigurierbaren Sicht (s. Abb. 1) präsentiert. Jede Ressource wird in einem eigenen Fenster angezeigt (*Panel*), eine Sicht (*View*) besteht aus einem oder mehreren Fenstern. Das System stellt dem Benutzer eine Reihe vorkonfigurierter Sichten zu Verfügung (Standardsicht, Korpussicht usw.). Darüber hinaus kann jeder registrierte Benutzer sich aus den bestehenden Ressourcen/Panels eine oder mehrere private Sichten erstellen und nutzen (mehr dazu in Klein & Geyken 2010: Abschnitt 6.4).

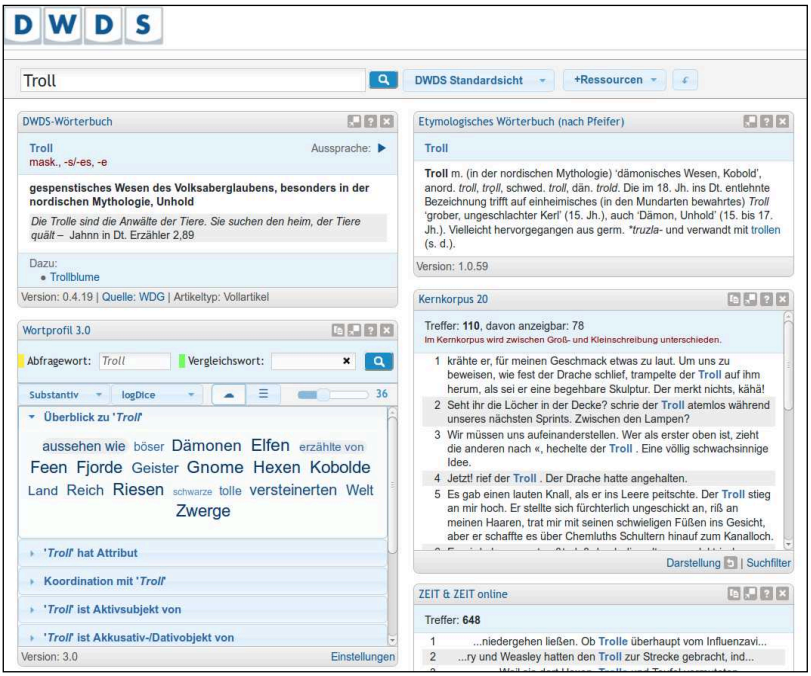


Abb. 1: Panelbasierte Sicht des DWDS, Stichwort ‚Troll'.

Nutzer können sich auf diese Weise schnell ein Bild über ein Wort (oder eine Wortkombination), seine Bedeutung und seinen Gebrauch in der geschriebenen Gegenwartssprache machen.

Der modulare Aufbau des Digitalen Lexikalischen Systems – Informationen aus heterogenen Ressourcen werden „unter dem Dach" einer Suchanfrage angeboten – erleichtert die Integration weiterer Ressourcen, sofern diese hinsichtlich Format und Annotation zu den vorhandenen Ressourcen kompatibel sind. Die vorhandenen Ressourcen wurden wie folgt aufbereitet:

- **Linguistische Annotation:** Die Korpora wurden mit den an der BBAW entwickelten sprachtechnologischen Werkzeugen linguistisch annotiert. Die Segmentierung der Texte, die Tokenisierung und die Wortartenannotation erfolgen mit dem Part-of-Speech-Tagger *moot* (Jurish, 2003). Für die Lemmatisierung der Textwörter wird die *TAGH*-Morphologie verwendet (Geyken & Hanneforth 2006). Dies ermöglicht die Formulierung komplexerer linguistischer Abfragen, z.B. Suchmuster als Kombinationen von Wortformen, Lemmata und Wortarten. Die für das DWDS entwickelte Suchmaschine *DDC* kann solche komplexen Abfragen bearbeiten. Auch komplexe statistische Auswertungen, z.B. zu typischen Wortkombi-

nationen in bestimmten syntaktischen Beziehungen, bauen auf diese Annotationen auf.

- **Korpusrepräsentation:** Korpora und lexikalische Ressourcen sind durchgängig in XML nach den Kodierungsregeln der Text Encoding Initiative (TEI-P5) ausgezeichnet. Dies umfasst sowohl die Auszeichnung der Primärdaten als auch die Bereitstellung der Metadaten.

Das DeRiK-Korpus soll konform zu den genannten Anforderungen aufbereitet werden. Hinsichtlich der linguistischen Annotation müssen die an der BBAW entwickelten sprachtechnologischen Verfahren an die sprachlichen Besonderheiten schriftlicher internetbasierter Kommunikation angepasst werden, um ein qualitativ akzeptables Analyseergebnis zu erzielen. Vor allem müssen diejenigen Phänomene berücksichtigt und angemessen behandelt werden, die charakteristisch für das interaktionsorientierte Schreiben in sozialen Medien sind, z.B. Schnellschreibphänomene und Phänomene geschriebener Umgangssprache sowie Elemente der „Netzsprache", die in redigierten Texten nur in Ausnahmefällen auftreten (z.B. Emoticons, Aktionswörter, Adressierungen). Für die Bearbeitung dieser Aufgabe wird derzeit in Kooperation mit dem DFG-Netzwerk *Empirikom* eine Community-Shared-Task zur automatischen linguistischen Analyse und Annotation deutscher IBK-Daten vorbereitet (vgl. http://empirikom.net/bin/view/Themen/SharedTask). Vorschläge für die Erweiterung des STTS-Tagsets um Einheiten für das POS-Tagging von „Netzsprache"-Phänomenen sind in Bartz et al. (2013) beschrieben und werden in der Arbeitsgruppe zur Erweiterung von STTS diskutiert (s. JLCL 2013, Heft 1).

Hinsichtlich der Repräsentation der Primärdaten müssen die DeRiK-Daten in einem TEI-kompatiblen Format repräsentiert werden. Für diesen Zweck wurde von 2010 bis 2012 ein auf den TEI-P5-Formaten basierendes – und deshalb TEI-konformes – und auf die Besonderheiten der schriftlichen internetbasierten Kommunikation angepasstes Repräsentationsschema entwickelt. Die Eckpunkte dieses Schemas sowie seiner geplanten Weiterentwicklung im Zusammenhang mit der Special Interest Group „Computer-Mediated Communication" der TEI werden im folgenden Abschnitt beschrieben.

5. TEI-Repräsentation der Korpusdaten: Stand der Arbeiten und Perspektiven

Da IBK-Korpora einen vergleichsweise neuen Typus von Korpora mit spezifischen strukturellen und linguistischen Besonderheiten darstellen (vgl. Storrer 2013a: Abschnitt 4), existieren in den 'Digital Humanities' bislang keine Standards oder Quasi-Standards für die Repräsentation der in ihnen erfassten Datentypen und Genres. Wer bislang annotierte Korpora mit Sprachdaten aus Genres internetbasierter Kommunikation aufbaut, muss dafür i.d.R. eigene Annotationsschemata entwickeln. So wurde beispielsweise für das Dortmunder Chat-Korpus ein Schema definiert, das speziell auf die Strukturmodellierung von Chat-Mitschnitten (Logfiles), die Annotation unterschiedlicher Typen von Nutzerbeiträgen sowie die Auszeichnung ausgewählter „Netzsprache"-Phänomene zugeschnitten ist (vgl. Beißwenger 2013).

Auch für DeRiK stellt sich das Problem fehlender Standards im Bereich der Repräsentation von Korpusdokumenten mit Sprachdaten aus IBK-Genres. Da die bereits in der DWDS-

Korpusinfrastruktur vorhandenen Ressourcen in TEI-P5 repräsentiert sind, sollen auch die DeRiK-Daten auf der Basis von TEI-P5 annotiert werden. Zum gegenwärtigen Zeitpunkt finden sich in den in TEI-P5 enthaltenen Formaten allerdings noch keine Modelle, die auf die Repräsentation der strukturellen und sprachlichen Besonderheiten von IBK-Genres zugeschnitten sind oder die sich ohne Weiteres für die Repräsentation von IBK-Daten übernehmen lassen. Allerdings bietet das Encoding Framework der TEI die Möglichkeit, vorhandene Modelle auf die Erfordernisse neuer, bislang noch nicht im Standard berücksichtigter Genres anzupassen (in der TEI-Terminologie als *customization* bezeichnet):
Because the TEI Guidelines must cover such a broad domain and user community, it is essential that they be customizable: both to permit the creation of manageable subsets that serve particular purposes, and also to permit usage in areas that the TEI has not yet envisioned. (http://www.tei-c.org/Guidelines/Customization/)
Bei der *customization* erlaubt ist u.a. das Hinzufügen neuer, die Entfernung nicht benötigter und die Umbenennung vorhandener TEI-Elemente, die Anpassung von Inhaltsmodellen, das Hinzufügen und Entfernen von Attributen sowie die Modifikation von Attributwertlisten zu Elementen. Solange dabei bestimmte Regeln für die Spezifikation neuer Elemente und Attribute und für die Dokumentation individuell vorgenommener Modifikationen eingehalten werden, bleiben auf diese Weise erzeugte Repräsentationsschemata – obwohl sie Modelle umfassen, die selbst nicht Teil des Standards sind – kompatibel zu TEI-P5.

Das für DeRiK entwickelte Schema macht von dieser Möglichkeit der Anpassung Gebrauch. Die Basis für das Schema bildet das Modul „Basic text structure“ aus TEI-P5; dieses Modul wird für die Zwecke der Annotation von IBK-Genres spezifisch erweitert und modifiziert. Zentrale Komponenten des Schemas sind

- eine Komponente für die Beschreibung der *Makrostruktur* von Genres schriftlicher internetbasierter Kommunikation;
- eine Komponente für die Beschreibung ausgewählter „netztypischer“ Stilelemente innerhalb einzelner Nutzerbeiträge (= *Mikrostruktur* von IBK).

Im Folgenden geben wir einen Überblick über einige wichtige Eckpunkte des Schemas; eine ausführliche Beschreibung findet sich in Beißwenger et al. (2012), das zugehörige ODD-Dokument sowie Annotationsbeispiele können unter http://empirikom.net/bin/view/Themen/CmcTEI heruntergeladen werden.

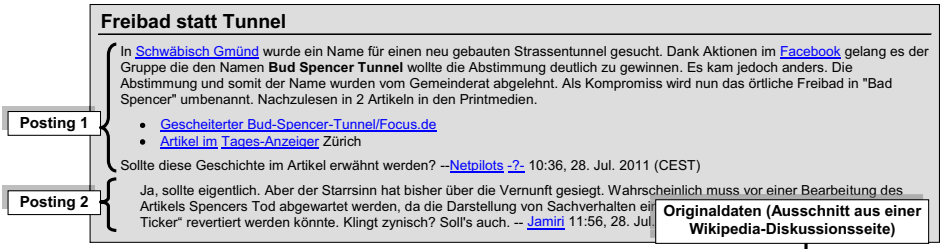
Die grundlegende Modellierungseinheit des Schemas auf der Ebene der *Makrostruktur* von IBK-Dokumenten bildet das *Posting*, das spezifiziert ist als eine Zeichensequenz, die zu einem bestimmten Zeitpunkt von einem Nutzer – etwa durch Betätigung der Eingabetaste – *en bloc* an den Server übermittelt und anschließend als neuer Nutzerbeitrag am Bildschirm angezeigt wird.

Postings werden in unterschiedlichen IBK-Genres auf unterschiedliche Arten zu größeren Einheiten zusammengeordnet. Das Schema unterscheidet zwischen zwei Typen von IBK-Makrostrukturen:

- dem Strukturtyp ‚Logfile’, bei dem Postings, auf- oder absteigend, linear chronologisch nach ihren Eintreffenszeitpunkten beim Server dargestellt werden; die Abfolge der Postings wird dabei vom Server festgelegt;

- dem Strukturtyp ‚Thread', bei dem Postings entlang einer *oben/unten*- und einer *links/rechts*-Dimension auf der Bildschirmseite platziert werden. *Oben/unten* symbolisiert dabei prototypischerweise eine *vorher/nachher*-Relation, unterschiedliche Grade der Rechtseinrückung auf der *links/rechts*-Dimension symbolisieren thematische Bezüge auf Vorgängerpostings. In manchen Systemen (z.B. klassischen Foren mit Baumstrukturdarstellung) wird die Platzierung auf der *oben/unten*- und auf der *links/rechts*-Dimension automatisch erzeugt; spezifiziert der Nutzer ein bestimmtes Vorgängerposting als Bezugsbeitrag für sein eigenes Posting, so wird das eigene Posting nach der Verschickung relativ zum Bezugsbeitrag um eine Ebene eingerückt. Auf Wikipedia-Diskussionsseiten können Nutzer die Platzierung ihrer Postings – sowohl in der Horizontalen wie auch in der Vertikalen – hingegen selbst frei festlegen.

Zu jedem Posting wird ein Nutzer als Autor spezifiziert, der Name des Nutzer wird dabei durch eine ID ersetzt. Die Modellierung der als Posting-Autoren im Dokument dokumentierten Nutzer inklusive der zu ihnen im Dokument gegebenen Informationen (z.B. Inhalt der individuellen Nutzersignatur) wird in Personenprofilen im Dokument-Header gespeichert. Über eine ID-Referenz im <posting>-Element kann die ID auf den tatsächlichen Nutzernamen bezogen werden. Für die Bereitstellung des Korpus lassen sich die Korpusdokumente durch die Trennung der Nutzerinformationen von den Postings einfach anonymisieren. Abb. 2 zeigt am Beispiel eines Ausschnitts aus einer Wikipedia-Diskussionsseite die Annotation zweier Postings. Die Nutzer sind im Element <listPerson> modelliert, die Postings sind über das Attribut *@who* den Einträgen in der Liste zugeordnet. Die Veröffentlichungszeitpunkte der Postings sind im Element <timeline> modelliert; die Postings sind über das *@synch*-Attribut den Einträgen in der Timeline zugeordnet.




```
<listPerson>
        <person xml:id="A01">
                <persName>Netpilots</persName>
                <signatureContent><ref target="http://de.wikipedia.org/wiki/Benutzer:Netpilots">Netpilots</ref><ref
                target="http://de.wikipedia.org/wiki/Benutzer_Diskussion:Netpilots">-Ì¡-</ref></signatureContent>
        </person>
        <person xml:id="A02">
                <persName>Jamiri</persName>
                <signatureContent><ref
                target="http://de.wikipedia.org/wiki/Benutzer:Jamiri">Jamiri</ref></signatureContent>
        </person>
        …
</listPerson>
<front>
        <timeline>
                <when xml:id="t01" absolute="2011-07-27T16:46:00"/>
                <when xml:id="t02" absolute="2011-07-28T10:36:00"/>
                …
        </timeline>
</front>
<body>
        <div type="thread">
                <head>Freibad statt Tunnel</head>
                <posting synch="#t01" who="#A07">
                        <p>In<ref target="http://de.wikipedia.org/wiki/Schw%C3%A4bisch_Gm%C3%BCnd">Schwäbisch
                        Gmünd</ref> wurde ein Name für einen neu gebauten Strassentunnel gesucht. Dank Aktionen im <ref
                        target="http://de.wikipedia.org/wiki/Facebook">Facebook</ref> gelang es der Gruppe die den Namen
                        Bud Spencer Tunnel wollte die Abstimmung deutlich zu gewinnen. Es kam jedoch anders. Die
                        Abstimmung und somit der Name wurden vom Gemeinderat abgelehnt. Als Kompromiss wird nun das
                        örtliche Freibad in "Bad Spencer" umbenannt. Nachzulesen in 2 Artikeln in den Printmedien.</p>
                        <list>
                                <item><ref target="http://www.focus.de/panorama/welt/stuermische-ratssitzung-kein-bud-
                                spencer-tunnel-in-schwaebisch-gmuend_aid_649932.html,">Gescheiterter Bud-Spencer-
                                Tunnel/Focus.de</ref></item>
                                <item><ref target="http://www.tagesanzeiger.ch/leben/gesellschaft/Grosse-Hysterie-um-einen-
                                alten-Mann-/story/17754241">Artikel im</ref> <ref target="http://de.wikipedia.org/wiki/Tages-
                                Anzeiger">Tages-Anzeiger</ref> Zürich</item>
                        </list>
                        <p>Sollte diese Geschichte im Artikel erwähnt werden? -- <autoSignature/></p>
                </posting>
                <posting synch="#t02" who="#A06" indentLevel="1">
                        <p>Ja, sollte eigentlich. Aber der Starrsinn hat bisher über die Vernunft gesiegt. Wahrscheinlich muss
                        vor einer Bearbeitung des Artikels Spencers Tod abgewartet werden, da die Darstellung von
                        Sachverhalten einer noch lebenden Person sonst als „Live-Ticker“ revertiert werden könnte. Klingt
                        zynisch? Soll's auch. -- <autoSignature/></p>
                </posting>
                …
        </div>
</body>
```

Abb. 2: Encoding-Beispiel: Wikipedia-Diskussionsseite.

Während auf der *Makroebene* einzelne Postings annotiert sowie Strukturen oberhalb der Postingebene modelliert werden, beschreiben wir auf der *Mikroebene* von IBK-Dokumenten

sprachliche Besonderheiten *innerhalb* von Postings. Die Annotation bezieht sich hier also auf die schriftlichen Beiträge der Verfasser zum Kommunikationsgeschehen, die als Postings an den Server geschickt werden. Von besonderem Interesse für die korpusgestützte Analyse und die lexikographische Bearbeitung internetbasierter Kommunikation sind dabei solche Einheiten, die als „typisch netzsprachlich" gelten:

- *Emoticons*, die durch die Kombination von Interpunktions-, Buchstaben- und Sonderzeichen gebildet werden, ikonisch fundiert sind und daher übereinzelsprachlich verwendet werden können. Emoticons dienen typischerweise der emotionalen Kommentierung, der Markierung von Ironie oder der Bewertung von Partneräußerungen. 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.
- *Aktionswörter* (*interaction words*), die zur sprachlichen Beschreibung von Gesten, mentalen Zuständen oder Handlungen verwendet werden und als Emotions- oder Illokutionsmarker, Ironiemarker oder für die spielerische Nachbildung fiktiver Handlungen verwendet werden. Sie sind einzelsprachlich gebunden und basieren auf einem Wort – typischerweise einem unflektierten Verbstamm –, das entweder alleine steht (*lach*, *freu*, *grübel*) oder um weitere Einheiten erweitert ist (*malnachdenk*, *stirnrunzel*).
- *Adressierungen*, mit denen Kommunikationsbeiträge an einen bestimmten anderen Kommunikationsbeteiligten oder eine Gruppe von Kommunikationsbeteiligten adressiert werden und die häufig zur Sicherstellung von thematischen Bezügen auf Vorbeiträge (des/der benannten Adressaten) verwendet werden (*@tinchen*, *@alle*).
- *„Interaction templates"*, die ähnliche Funktionen übernehmen wie Emoticons und Aktionswörter, bei denen es sich aber weder um tastaturschriftlich erzeugte noch um frei formulierte sprachliche Einheiten handelt, sondern um Einheiten, die – als statische oder animierte Grafiken oder als Kombinationen aus Grafiken und vordefinierten Textbausteinen – von den Nutzern durch Eingabe bestimmter Codes oder per Auswahl aus einem Grafikmenü in ihre Beiträge eingefügt werden. Beispiele sind die sog. „Grafik-Smileys" sowie Vorlagenbausteine, die in der Wikipedia durch Eingabe bestimmter Codes in Beiträge auf Diskussionsseiten integriert werden können (z.B. im Rahmen von Abstimmungen, vgl. die beiden Beispiele in Abb. 3).

Die vier Typen von Einheiten haben gemeinsam, dass sie typischerweise nicht syntaktisch integriert sind und sowohl im linken oder rechten Außenfeld von Sätzen auftreten wie auch an nahezu beliebiger Position in Form von Parenthesen eingeschoben werden können. Funktional sind sie spezialisiert auf Aufgaben im Bereich der Handlungskoordination im Dialog, der emotionalen Kommentierung und der Respondierung vorangegangener Partneräußerungen. Sie ähneln damit dabei den Interjektionen beziehungsweise den Einheiten, die die

„Grammatik der deutschen Sprache“ (GDS, Zifonun et al. 1997, online: GRAMMIS, s. http://hypermedia.ids-mannheim.de/) als *Interaktive Einheiten* beschreibt und zu denen neben den Interjektionen auch die Responsive (*ja*, *nein*) gehören.

Das DeRiK-Schema führt für die oben beschriebenen Einheiten jeweils eigene Annotationskategorien ein, die in Anlehnung an die Kategorie der GDS einer gemeinsamen Oberkategorie „interaction sign“ zugeordnet sind und damit als IBK-spezifische Erweiterungen der Kategorie der interaktiven Einheiten dargestellt werden. Jeder Einheitentyp ist durch ein eigenes XML-Element beschrieben und kann durch eine Reihe von Attributen subklassifiziert werden. Die Schemakomponente für die Beschreibung der „interaction signs“ ist im Detail in Beißwenger et al. (2012: Abschnitt 3.5.1) beschrieben und begründet. Eine Übersicht über die Kategorie der „interaction signs“ gibt Abb. 3.

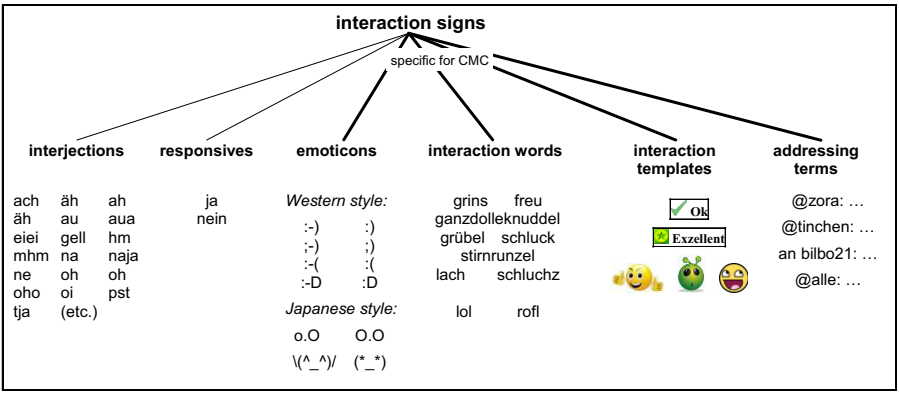


Abb. 3: „Interaction signs“.

Das für DeRiK entwickelte TEI-Schema bildet eine der Grundlagen für die Arbeit in der 2013 neu eingerichteten Special Interest Group (SIG) „Computer-Mediated Communication“ der *Text Encoding Initiative* (http://www.tei-c.org/Activities/SIG/CMC/). In der Gruppe arbeiten Vertreterinnen und Vertreter von Korpusprojekten zu verschiedenen europäischen Sprachen – darunter auch die DeRiK-Beteiligten – gemeinsam an der Erarbeitung von Vorschlägen für die Integration von Modellen für die Annotation von IBK-Genres in das Encoding Framework der TEI. Dabei wird u.a. angestrebt, eine Basisstruktur für die Repräsentation von IBK-Korpora zu entwickeln, die auch die Darstellung der Kommunikation in multimodalen Online-Umgebungen erlaubt, sowie ein Schema für die Repräsentation von Metadaten zu Sprachdaten aus Online-Umgebungen zu entwickeln. Das DeRiK-TEI-Schema soll, wo erforderlich, an die von der SIG entwickelten Vorschläge angepasst werden, um die DeRiK-Daten interoperabel zu den Korpusdaten der übrigen in der SIG beteiligten Projekte zu repräsentieren.

6. Anwendungsszenarien: Korpusgestützte Sprachanalyse und gegenwartssprachliche Lexikographie

Für den Bereich der korpusgestützten linguistischen Sprachanalyse wird das DeRiK-Korpus neue Möglichkeiten eröffnen, die Sprachverwendung in sozialen, internetbasierten Medien sowie deren Auswirkungen auf die geschriebene deutsche Gegenwartssprache empirisch zu untersuchen. Durch seine Integration in die Korpusinfrastruktur des DWDS wird es u.a. möglich,

- auf Basis eines Referenzkorpus typische Stilmerkmale des interaktionsorientierten Schreibens im Netz (z.B. Schnellschreibphänomene, Phänomene der konzeptionellen Mündlichkeit, Emoticons und Aktionswörter) für unterschiedliche Genres und Kontexte internetbasierter Kommunikation qualitativ und quantitativ zu untersuchen;
- die Variationsbreite bei der Verwendung dieser Stilmerkmale innerhalb der internetbasierten Kommunikation empirisch darzustellen und für diesen Bereich Faktoren sprachlicher Variation unter den Bedingungen technischer Vermittlung herauszuarbeiten;
- das Auftreten dieser Phänomene in IBK-Genres und in der redigierten Schriftlichkeit außerhalb des Netzes vergleichend zu untersuchen;
- den Wandel der Sprachverwendung innerhalb der internetbasierten Kommunikation zu untersuchen;
- durch vergleichende Untersuchungen zu sprachlichen Merkmalen in IBK-Daten und in den DWDS-Textkorpora charakteristische Unterschiede des *text-* und des *interaktionsorientierten Schreibens* herauszuarbeiten und für didaktische Zwecke (z.B. die Vermittlung im Deutschunterricht) fruchtbar zu machen.[3]

Darüber hinaus kann das Korpus auch als Ressource für die Sprachdidaktik genutzt werden – beispielsweise bei der Entwicklung didaktischer Materialien für den Bereich „Reflexion über Sprache“ des sprachbezogenen Deutschunterrichts oder dazu, im Modus des „Forschenden Lernens“ mit Schülerinnen und Schülern Besonderheiten der Sprachverwendung und der sprachlichen Variation in sozialen Medien anhand der DWDS-Korpora (inkl. DeRiK) selbst zu erarbeiten.

Da die Ressourcen des DWDS-Projekts an der BBAW in erster Linie als Basis für die Aktualisierung eines gegenwartssprachlichen Wörterbuches verwendet werden (vgl. Abschnitt 4), wird DeRiK darüber hinaus als Ressource für die gegenwartssprachliche Lexikographie Verwendung finden. Durch die Miteinbeziehung des IBK-Korpus wird es möglich, Phänomene des lexikalischen Wandels im Gegenwartsdeutschen, die auf die Sprachverwendung im Netz zurückzuführen sind, korpusgestützt lexikographisch zu bearbeiten. Im Folgenden präsentieren wir zwei Beispiele für sprachliche Zeichen, die, zumindest in einer Bedeutung, der Domäne der internetbasierten Kommunikation entstammen und für die eine angemessene lexikographische Beschreibung daher ohne die Konsultation von Sprachdaten

3 Zur Unterscheidung zwischen text- und interaktionsorientiertem Schreiben und ihrer didaktischen Relevanz für die Bewertung der schriftlichen Sprachverwendung in der internetbasierten Kommunikation vgl. Storrer (2012, 2013).

aus IBK-Genres nicht gelingen kann. Die Beispiele veranschaulichen die Notwendigkeit der Einbeziehung von IBK-Ressourcen für eine umfassende gegenwartssprachliche Lexikographie.

‚Troll'

Zum Stichwort *Troll* verzeichnet das auf dem „Wörterbuch der deutschen Gegenwartssprache" basierende DWDS-Wörterbuch eine Bedeutung (s. auch Abb. 1):

„gespenstisches Wesen des Volksaberglaubens, besonders in der nordischen Mythologie, Unhold"

Der Grund dafür, dass dieses Wort als Bezeichnung für eine 'Person, die in internetbasierten Kommunikationsumgebungen eine störende Rolle einnimmt, meist in provokativer Absicht', nicht in diesem Wörterbuch verzeichnet ist, ist der Zustand des WDG, das den Sprachgebrauch bis etwa Mitte der siebziger Jahre verzeichnet und zurzeit aktualisiert wird. Aber auch die Korpora des DWDS (die jüngsten Texte dort sind aus dem Jahr 2010) erzählen keine andere Geschichte. Der ‚Troll' (und das ‚Trollen') als – von den meisten als lästig empfundener – Teil der Netzkultur kommt in diesen Korpora (noch nicht) vor. Umso wichtiger ist es, die Möglichkeit zur Recherche in einem aktuellen Korpus der internetbasierten Kommunikation zu haben, spätestens dann, wenn der Artikel „Troll" im DWDS-Wörterbuch überarbeitet werden soll.

‚lol'

Im Jahr 2011 wurde das Akronym 'lol' für 'laughing out loud', ein in der internetbasierten Kommunikation häufig verwendetes Aktionswort, durch Aufnahme in das „Oxford English Dictionary" geadelt. Das OED erkennt damit die lexikalische Produktivität der netzbasierten Kommunikation an.

Im DWDS-Wörterbuch findet sich hierzu, aus den oben bereits genannten Gründen, nichts. Die Korpora des DWDS sind da etwas ergiebiger. Allein 46 Treffer findet man in der ZEIT. In den meisten dieser Belege sehen es die Autoren allerdings als notwendig an, das Akronym aufzulösen – vertrauen also nicht auf sofortiges Verstehen des Kürzels bei ihrer Leserschaft.

In der Jugendsprache geht diese Abkürzung neuerdings eigene Wege. Immer öfter vernimmt man die Adjektivbildung 'lollig', die aus 'lol' gebildet wurde und laut Auskunft einiger Sprecher eindeutig positiv-anerkennend gemeint ist (etwa: „echt witzig").

Wortneuschöpfungen im Bereich gruppenspezifischer Sprachen, wie etwa der Jugendsprache, sind oft Gelegenheitsbildungen. Ihre Verwendung wird deshalb von Lexikographen erst einmal längere Zeit beobachtet, ehe über eine (Nicht-)Aufnahme ins Wörterbuch entschieden wird. Stellt man dann fest, dass eine Neuschöpfung sich in der Netzkultur oder auch darüber hinaus durchsetzen konnte, wird für die Aufnahme dieses Wortes entschieden. Meist sind dann auch die Verwendungsregeln und -formen so stabil, dass eine verlässliche lexikographische Beschreibung möglich ist. Die Bedeutung(en) und die unterschiedlichen kommunikativen Funktionen kann man aber nur nachvollziehen und angemessen beschreiben, wenn man ein Korpus internetbasierter Kommunikation mit einer gewissen historischen Tiefe zur Verfügung hat. Deshalb sind die Nachhaltigkeit des Projektes, die eine dauerhafte Weiterentwicklung des Korpus ermöglicht, und die Interoperabilität der Daten, die ihre

Verknüpfung mit anderen Korpora gewährleistet, von zentraler Bedeutung für die lexikographische Anwendung.

7. Ausblick

IBK-Korpora stellen einen Korpustyp neuer Art dar, zu dem im Bereich der 'Digital Humanities' noch in verschiedener Hinsicht Forschungsbedarf besteht: Diverse grundlegende Fragen hinsichtlich des Aufbaus, der Annotation und der Integration in Korpusinfrastrukturen sind für diesen Korpustyp noch nicht abschließend geklärt (vgl. Storrer 2013a: Abschnitt 4).

DeRiK versteht sich deshalb als ein Projekt, dessen primäres Ziel zwar der Aufbau und die Bereitstellung eines Korpus ist, das aber zugleich auch den Rahmen dafür bildet, methodische Fragen in Bezug auf die Integration und Beschreibung von IBK-Daten in linguistische Korpora zu sondieren und Lösungsvorschläge dafür zu erarbeiten. Zentrale Fragen sind u.a. die angemessene und an Standards (z.B. TEI) orientierte Modellierung der Primärdaten, die Repräsentation von Metadaten zu IBK-Genres sowie Fragen der linguistischen Annotation und der Anpassung von Werkzeugen für die automatische Sprachverarbeitung. Entscheidend für den Erfolg der Unternehmung ist die enge Zusammenarbeit mit Initiativen und Projekten, die sich im nationalen und internationalen Rahmen mit diesen Fragen befassen – als Beispiele seien hier das schon erwähnte DFG-Netzwerk „Empirische Erforschung internetbasierter Kommunikation", die Special Interest Group „Computer-Mediated Communication" im Rahmen der TEI sowie die in Vorbereitung befindliche Shared Task zur linguistischen Annotation deutschsprachiger IBK-Daten genannt. Von der engen Zusammenarbeit mit Infrastrukturprojekten im Bereich der Digital Humanities wie CLARIN (http://www.clarin.eu) und, national, CLARIN-D (http://de.clarin.eu) erwarten wir Fortschritte in Richtung auf eine umfassende Interoperabilität der entstehenden Ressourcen. Im Rahmen der Special Interest Group sollte es gelingen, ein Referenzschema zu entwickeln, das als Grundlage für die Annotation im Aufbau befindlicher Korpora zu verschiedenen europäischen Sprachen und unterschiedlichen IBK-Genres dienen kann und das damit die Interoperabilität und Verknüpfbarkeit dieser Ressourcen ermöglicht – die damit für die korpusgestützte Bearbeitung auch von sprachübergreifenden Forschungsfragen von Nutzen sein können. CLARIN ist auch ein geeigneter Rahmen, um das entstehende Korpus a) für die interessierten Fachcommunitys sichtbar zu machen – hierfür bietet das „Virtual Language Observatory"[4] eine geeignete Plattform – und b) die Daten zu distribuieren. Dadurch, dass nur Texte akquiriert werden, deren Lizenzstatus eine Redistribuierung der Daten (und Annotationen) erlauben, können interessierte Nutzer nicht nur über die DWDS-Webseite in den Daten recherchieren, sondern sich das gesamte Korpus oder Teile davon herunterladen und in ihrer individuellen Arbeitsumgebung nutzen.

Aus lexikographischer Perspektive ist die kontinuierliche Weiterentwicklung von IBK-Korpora von besonderer Bedeutung, damit Prozesse der Lexikalisierung und des Bedeutungswandels auch in diesem Bereich erfasst und beschrieben werden können. Umso wichtiger ist es, dass für DeRiK grundlegende Fragen der Modellierung und Annotation von

4 S. http://www.clarin.eu/vlo.

IBK-Daten bereits in einer frühen Projektphase geklärt werden. Aktuelle Arbeitsschwerpunkte liegen daher zum einen auf der Entwicklung von Standardisierungsvorschlägen im Kontext der TEI und zum anderen auf der Anpassung von Sprachverarbeitungswerkzeugen. Daneben ist für 2014 die Erhebung einer ersten Tranche von Sprachdaten aus verschiedenen, rechtlich unproblematischen Online-Umgebungen geplant.

8. Literatur

Bartz, T.; Beißwenger, M. & Storrer, A. (2013, im Erscheinen). Optimierung des Stuttgart-Tübingen-Tagset für die linguistische Annotation von Korpora zur internetbasierten Kommunikation: Phänomene, Herausforderungen, Erweiterungsvorschläge. In: *Journal for Language Technology and Computational Linguistics (Themenheft „Das STTS-Tagset für Wortartentagging – Stand und Perspektiven").*

Beißwenger, M. (Hrsg., 2001). *Chat-Kommunikation. Sprache, Interaktion, Sozialität & Identität in synchroner computervermittelter Kommunikation. Perspektiven auf ein interdisziplinäres Forschungsfeld.* Stuttgart: ibidem.

Beißwenger, M. (2013). Das Dortmunder Chat-Korpus. In: *Zeitschrift für germanistische Linguistik* 41/1, pp. 161-164.

Beißwenger, M. & Storrer, A. (2008). Corpora of Computer-Mediated Communication. In Lüdeling, A. & Kytö, M. (eds), *Corpus Linguistics*. *An International Handbook,* vol. 1. Berlin, de Gruyter, pp. 292-308.

Beißwenger, M., Ermakova, M., Geyken, A., Lemnitzer, L. & Storrer, A. (2012). A TEI schema for the Representation of the Computer-mediated Communication. In: *Journal of the Text Encoding Initiative* 3. http://jtei.revues.org/476 (DOI: 10.4000/jtei.476).

Busemann, K. (2013): Wer nutzt was im Social Web? Ergebnisse der ARD/ZDF-Onlinestudie 2013. Media Perspektiven 7-8/2013, 373–385. http://www.ard-zdf-onlinestudie.de/fileadmin/Onlinestudie/PDF/Busemann.pdf

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

Crystal, D. (2011). *Internet Linguistics. A Student Guide.* New York: Routledge.

Forsyth, E. N. & Martell, C. H. (2007). Lexical and Discourse Analysis of Online Chat Dialog. In *Proceedings of the First IEEE International Conference on Semantic Computing (ICSC 2007)*, pp. 19-26.

Geyken, A. (2007). The DWDS corpus: A reference corpus for the German language of the 20th century. In: Fellbaum, Ch. (ed.), *Collocations and Idioms*. London: continuum, pp. 23-40.

Geyken A. & Hanneforth T. (2006). TAGH: A Complete Morphology for German based on Weighted Finite State Automata. In: Yli-Jyrä A, Karttunen L, Karhumäki J, (eds.), *Finite State Methods and Natural Language Processing.* Berlin/Heidelberg: Springer, 55-66.

Grimm, J. & Grimm W. (1852-1971). *Deutsches Wörterbuch. Erstbearbeitung*, 33 Bände, Leipzig:Hirzel Verlag

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

Herring, S. C. (ed., 2010/2011). Computer-Mediated Conversation. *Special Issue of Language@Internet,* vol. 7/8. http://www.languageatinternet.org/articles/2010, http://www.languageatinternet.org /articles/2011

Jurish, B. (2003). *A Hybrid Approach to Part-of-Speech Tagging. Final report, project 'Kollokationen im Wörterbuch'.* Berlin: BBAW; 2003. http://www.dwds.de/dokumentation/tagger/.

Kestemont, M., Peersman, C., De Decker, B., De Pauw, G., Luyckx, K., Morante, R. Vaassen, F., van de Loo, J., Daelemans, 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, pp. 1569-1572.

King, B. W. (2009). Building and Analysing Corpora of Computer-Mediated Communication. In: Baker, P. (ed.), *Contemporary corpus linguistics*. London: Continuum, pp. 301-320.

Klappenbach, R. & Steinitz, W. (eds., 1964-1977). *Wörterbuch der deutschen Gegenwartssprache (WDG).* 6 Bände. Berlin: Akademie-Verlag.

Klein, W., Geyken, A. (2010). Das Digitale Wörterbuch der Deutschen Sprache (DWDS). In Heid, U., Schierholz, S., Schweickard, W., Wiegand, H. E., Gouws, R. H. & Wolski, W. (eds), *Lexicographica*. pp. 79-96.

Koch, P. & Oesterreicher, W. (1994). Schriftlichkeit und Sprache. In: Günther, Hartmut/ Ludwig, Otto (eds.), *Schrift und Schriftlichkeit. Ein interdisziplinäres Handbuch internationaler Forschung.* Bd. 1. Berlin u.a., 587-604.

Oostdijk, N., M. Reynaert, V. Hoste & I. Schuurman (2013). The construction of a 500-million-word reference corpus of contemporary written Dutch. In J. Odijk & P. Spyns (eds.), *Essential Speech and Language Technology for Dutch.* Springer.

Openthesaurus. http://www.opentesaurus.org.

Pfeifer, W. (1993). *Etymologisches Wörterbuch des Deutschen*. Berlin:Akademie-Verlag, 2. Aufl.

Reffay, C., Betbeder, M.-L. & Chanier, T. (2012). Multimodal Learning and Teaching Corpora Exchange: Lessons learned in 5 years by the Mulce project. In: *International Journal of Technology Enhanced Learning (IJTEL)* 4, vol. 1/2. DOI: 10.1504/IJTEL.2012.048310

Reynaert, N., Oostdijk, O., De Clercq, O., van den Heuvel, H. & de Jong, F. (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, pp. 2693-2698.

Runkehl, K., Siever, T. & Schlobinski, P. (1998). *Sprache und Kommunikation im Internet. Überblick und Analysen.* Opladen: Westdeutscher Verlag.

Storrer, A. (2012). Neue Text- und Schreibformen im Internet: Das Beispiel Wikipedia. In: Juliane Köster & Helmuth Feilke (Hrsg.): *Textkompetenzen für die Sekundarstufe II.* Freiburg: Fillibach, 277-304.

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

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

[TEI-P5] TEI Consortium (ed., 2007). *TEI P5: Guidelines for Electronic Text Encoding and Interchange.* http://www.tei-c.org/Guidelines/P5/ (accessed 30 April 2013).

van Eimeren, B. & Frees, B. (2013): Rasanter Anstieg des Internetkonsums – Onliner fast drei Stunden täglich im Netz. Ergebnisse der ARD/ZDF-Onlinestudie 2013. In: Media Perspektiven 7-8/2013, 358–372. http://www.ard-zdf-onlinestudie.de/fileadmin/Onlinestudie/PDF/Eimeren_Frees.pdf

Zifonun, G., Hoffmann, L. & Strecker, B. (1997): *Grammatik der deutschen Sprache.* 3 Bände. Berlin/ New York: de Gruyter.