Michael Beißwenger, Harald Lungen, Jan Schallaböck,
John H. Weitzmann, Axel Herold, Pawel Kamocki,
Angelika Storrer und Julia Wildgans

Rechtliche Bedingungen für die Bereitstellung eines Chat-Korpus in CLARIN-D: Ergebnisse eines Rechtsgutachtens

1 Einleitung

Korpora zur Sprachverwendung in der internetbasierten Kommunikation, die frei für Forschung, Lehre und Sprachdokumentation zur Verfügung stehen, sind in der deutschen und internationalen Sprachressourcenlandschaft nur sehr begrenzt vorhanden. Zugleich besteht in den Bereichen Linguistik, Sprachvermittlung und Sprachtechnologie – allgemeiner in allen Bereichen der digitalen Geistes- und Sozialwissenschaften, die sich empirisch und sprachdatenbasiert mit Tendenzen der deutschen Gegenwartssprache und mit Phänomenen im Social Web beschäftigen – ein großer Bedarf an hochwertigen, für Recherchezwecke aufbereiteten Sprachressourcen. Wesentliche Gründe für die geringe Zahl an frei verfügbaren Ressourcen liegen im Fehlen von Standards für die Erhebung, Dokumentation, Repräsentation und Annotation von Sprachdaten aus Genres wie Chats, Online-Foren, Newsgroups, sozialen Netzwerken, Tweets, Weblogs oder mobilen Messaging-Anwendungen wie SMS und WhatsApp. Auch sind Anwendungen im Bereich des Natural Language Processing (NLP) bislang nur begrenzt in der Lage, Daten dieses Typs zu verarbeiten und automatisch mit linguistischen Annotationen zu versehen.

In den letzten Jahren gibt es zu verschiedenen dieser Desiderate eine rege, interdisziplinär ausgerichtete Forschung, die sich mit der Entwicklung von Standards für die Repräsentation und Annotation und mit der Anpassung von sprachtechnologischen Verfahren für den Bereich der internetbasierten Kommunikation bzw. den Sprachgebrauch in sozialen Medien befasst.¹ Lösungsvor-

¹ Das DFG-Netzwerk *Empirikom* hat in verschiedener Hinsicht einen Beitrag dazu geleistet, Lösungen für deutschsprachige Korpora zu entwickeln, u.a. durch die Initiierung einer Special Interest Group zum Thema in der Text Encoding Initiative (TEI, <http://tei-c.org>) sowie durch die Konzeption und Durchführung einer computerlinguistischen Community Shared Task zur automatischen linguistischen Annotation von Sprachdaten internetbasierter Kommunikation

schläge und Best Practices, die gegenwärtig in verschiedenen Projekten und für verschiedene Sprachen erarbeitet werden, lassen erwarten, dass sich die Rahmenbedingungen für die Erhebung, Repräsentation und linguistische Annotation von Korpora internetbasierter Kommunikation in den nächsten Jahren sichtbar verbessern werden.

Ein Desiderat, zu dem nach wie vor wenig gesichertes Wissen und keine Standards existieren, betrifft die Frage nach dem rechtlichen Status von Sprachdaten aus internetbasierten Kommunikationsumgebungen und nach den rechtlichen Rahmenbedingungen für deren Erhebung, Speicherung, Bearbeitung (Aufbereitung, Annotation) und Wiederbereitstellung in Form frei zugänglicher Korpusressourcen. Für die Klärung dieser Fragen fehlt Linguisten, Korpus- und Computerlinguisten die erforderliche juristische Expertise; zugleich sind Juristen typischerweise nicht am Aufbau von Sprachkorpora beteiligt.

Im vorliegenden Beitrag berichten wir über die Ergebnisse eines Rechtsgutachtens, das im Zusammenhang mit der Integration eines existierenden Korpus zur deutschsprachigen Chat-Kommunikation – des *Dortmunder Chat-Korpus* – in die Korpusinfrastrukturen von CLARIN-D durch die Anwaltskanzlei *iRights.law* erarbeitet wurde. Das vom Bundesministerium für Bildung und Forschung (BMBF) geförderte Verbundprojekt CLARIN-D² zielt auf den Aufbau und die Verstetigung einer digitalen Forschungsinfrastruktur für Sprachressourcen in den Geistes- und Sozialwissenschaften. Das Rückgrat der Infrastruktur bildet ein Verbund von *Zentren*, zu denen u.a. die Berlin-Brandenburgische Akademie der Wissenschaften (BBAW) und das Institut für Deutsche Sprache (IDS) in Mannheim gehören. CLARIN-D ist das deutsche Partnerprojekt der europäischen CLARIN-Initiative (*European Research Infrastructure for Language Resources and Technology*³), an dem 19 Länder mit Partnerprojekten beteiligt sind.

Die Integration des Dortmunder Chat-Korpus in die CLARIN-D-Infrastrukturen war Gegenstand eines sog. Kurationsprojekts, das aus Mitteln des Verbundprojekts finanziert wurde und in dem am Beispiel des Chat-Korpus Lösungen für die Erweiterung bestehender Korpusinfrastrukturen zum Deutschen um Korpusressourcen zur internetbasierten Kommunikation erarbeitet wurden. Eine zentrale Voraussetzung für die Arbeiten im Projekt bildete die Klärung der

(<http://sites.google.com/site/empirist2015/>). Einen aktuellen Überblick über Entwicklungen und Projekte zum Aufbau und zur Analyse von Korpora internetbasierter Kommunikation für diverse Genres und für verschiedene europäischen Sprachen bietet der Konferenzband Fišer & Beißwenger (2016).

² <http://www.clarin-d.de>

³ <https://www.clarin.eu/>

Frage, ob bzw. unter welchen Bedingungen eine Aufnahme des Chat-Korpus in CLARIN-D rechtlich unbedenklich ist.

In ihrem Gutachten beurteilten die beiden Gutachter das Korpus unter datenschutzrechtlicher, persönlichkeitsrechtlicher, urheber- und leistungsschutzrechtlicher Perspektive. Die dazu formulierten Einschätzungen beziehen sich zwar spezifisch auf das Dortmunder Chat-Korpus – d.h. auf eine Korpusresource, die bereits vollständig erhoben vorlag und nachträglich in die CLARIN-D-Infrastrukturen integriert werden sollte –, die im Gutachten formulierten grundsätzlichen Erwägungen sind aber über den konkreten Fall hinausgehend interessant, um ein Bild von der komplexen rechtlichen Situation in Bezug auf die Bereitstellung von Sprachdaten aus Genres internetbasierter Kommunikation in Korpora zu gewinnen.

Der vorliegende Beitrag gibt zunächst einen Überblick über Ziele, Arbeitspakete und Ergebnisse des CLARIN-D-Kurationsprojekts (Abschnitt 2) und anschließend eine ausführliche Zusammenfassung der wesentlichen Ergebnisse des Rechtsgutachtens in allgemeinverständlicher Form (Abschnitt 3). In Abschnitt 4 wird dargestellt, wie im Projekt ausgehend von den Empfehlungen des Rechtsgutachtens die Korpusdaten für die Integration und Bereitstellung anonymisiert und lizenziert wurden.

2 Überblick über das CLARIN-D-Kurationsprojekt *ChatCorpus2CLARIN*

Den Hintergrund für das in diesem Beitrag beschriebene Rechtsgutachten bildet das Projekt *ChatCorpus2CLARIN*, das im Zeitraum Mai 2015 bis Februar 2016 unter Leitung von Michael Beißwenger (Dortmund) und Angelika Storrer (Mannheim) an der Universität Mannheim durchgeführt wurde. Ziel des Projekts war es, die Daten eines existierenden Korpus internetbasierter Kommunikation (des *Dortmunder Chat-Korpus*, Beißwenger 2013), für die CLARIN-D-Infrastruktur aufzubereiten, an aktuelle Annotations- und Repräsentationsstandards im Bereich der Digital Humanities anzupassen und über die Korpuschnittstellen der CLARIN-Zentren am Institut für Deutsche Sprache (IDS) in Mannheim und an der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW) zugänglich zu machen. Das Projekt wurde als Kurationsprojekt im Rahmen von CLARIN-D vom Bundesministerium für Bildung und Forschung (BMBF) gefördert. Als zentrale Projektverantwortliche an den beiden CLARIN-D-Zentren wirkten für die BBAW Axel Herold und für das IDS Harald Längen mit. Als Projekt-

mitarbeiter war Eric Ehrhardt (Universität Mannheim) beteiligt. Daneben brachten Thomas Schmidt und Swantje Westpfahl (IDS, Abt. Pragmatik) als Kooperationspartner Expertise an der Schnittstelle zu Korpora gesprochener Sprache (*Forschungs- und Lehrkorpora gesprochenes Deutsch, FOLK*) in das Projekt ein.

Durch eine Remodellierung des Korpus auf Basis der Formate der Text Encoding Initiative (TEI), die Anreicherung um zusätzliche morphosyntaktische Annotationen sowie die Integration in CLARIN-D sollte der Zugriff auf die Ressource verbessert und sollten Nutzerinnen und Nutzern aus dem Bereich der germanistischen Linguistik und Computerlinguistik sowie aus dem Bildungsbereich (Sprachdidaktik, Lehramt Deutsch) erweiterte Recherche- und Analysemöglichkeiten an die Hand gegeben werden. Die Herstellung von Kompatibilität mit Standardformaten sollte die Interoperabilität der Ressource mit anderen Sprachressourcen verbessern; die Anpassung einer Annotationsumgebung sollte die Möglichkeiten des Einsatzes in der Lehre erweitern und es Nutzerinnen und Nutzern aus den genannten Bereichen ermöglichen, die in der Ressource vorhandenen Annotationen individuell anzupassen.

Das Projekt konnte auf die folgenden Vorarbeiten aufbauen:

- *Integration von Korpusressourcen in die CLARIN-D-Infrastrukturen:* Als Zentren verfügen die Arbeitsgruppenmitglieder am IDS und an der BBAW über die Expertise und die benötigten Technologien für die Integration neuer Sprachressourcen in die CLARIN-D-Infrastrukturen. Diese Voraussetzungen wurden für die Ausarbeitung der Zielformate (TEI, Repräsentation von morphosyntaktischen Informationen), für die Konvertierung der Ressource, für die Langzeitarchivierung und für die Integration in die Korpuschnittstellen am IDS und an der BBAW genutzt.
- *TEI-Modellierung:* Die Remodellierung der Ressource in einem TEI-Format konnte auf Vorarbeiten zur TEI-Repräsentation von Korpora in CLARIN-D-Infrastrukturen (Lüngen/Sperberg-McQueen 2012) sowie von Genres internetbasierter Kommunikation aufbauen, die von den Arbeitsgruppenmitgliedern im Rahmen des Korpusvorhabens DeRiK und des DFG-Netzwerks Empirikom sowie in Kooperation mit internationalen Partnern im Rahmen der Special Interest Group „Computer-Mediated Communication“ der TEI-Initiative (TEI-CMCSIG) durchgeführt wurden (vgl. Beißwenger et al. 2012, Chanier et al. 2014, Margaretha/Lüngen 2014).
- *Morphosyntaktische Annotation:* Für die Anreicherung der Ressource mit zusätzlichen morphosyntaktischen Annotationen (Parts of speech) konnte auf Ergebnisse aus der CLARIN-D-Initiative zur Erweiterung des „Stuttgart-Tübingen-Tagsets“ STTS (Schiller et al. 1999) um Kategorien u.a. für Genres „nichtstandardisierter Schriftlichkeit“ zurückgegriffen werden (vgl. Zins-

meister et al. 2014, speziell zur Anpassung von STTS an die Besonderheiten internetbasierter Kommunikation Bartz et al. 2014). Als Ziel-Tagset konnte im Projekt das erweiterte STTS-Tagset aus *EmpiriST2015*, einer von der Gesellschaft für Sprachtechnologie und Computerlinguistik (GSCL) geförderten Shared Task zur automatischen linguistischen Verarbeitung internetbasierter Kommunikation (Beißwenger et al. 2016), genutzt werden, das eine Weiterentwicklung der in der STTS-Arbeitsgruppe diskutierten Tagset-Entwürfe darstellt („STTS_IBK“, Beißwenger et al. 2015). In das Ziel-Tagset ist maßgeblich auch Expertise aus dem FOLK-Projekt der Abteilung Pragmatik am IDS eingeflossen, deren Mitarbeiter ebenfalls – für die morphosyntaktische Annotation gesprochener Sprache – an der CLARIN-D-Arbeitsgruppe zur Erweiterung von STTS beteiligt waren (vgl. Westpfahl 2014, Westpfahl/Schmidt 2016). Für die automatische Part-of-speech-Annotation konnten Werkzeuge aus einer Kooperation des IDS mit der Universität des Saarlandes im Rahmen des BMBF-Projekts www.schreibgebrauch.de genutzt werden, die bereits auf die Verarbeitung von Korpora internetbasierter Kommunikation angepasst waren (Horbach et al. 2014).

- *Rechtsgutachten*: Der Auftrag für die Erstellung des Rechtsgutachtens zu den Bedingungen der Integration der Ressource wurde auf Vorschlag des CLARIN-D LegalHelpDesk an die Anwaltskanzlei *iRights.law* erteilt. Die Kanzlei *iRights.law* (John Weitzmann u.a.) ist spezialisiert auf Rechtsfragen im Zusammenhang mit digitalen Medien und dem Internet und kooperiert bereits seit einigen Jahren mit CLARIN-D, beispielsweise bei der Erstellung der DFG-Handreichung zu rechtlichen Aspekten bei der Handhabung von Sprachkorpora (2015).

Auf dieser Grundlage wurden die folgenden wesentlichen Ergebnisse erzielt:

- *Repräsentation der Ressource in einem TEI-Format*: Das komplette Korpus wurde in TEI remodelliert. Das dafür im Projekt entwickelte Repräsentationsschema („CLARIN-D TEI-CMC“) erfasst sämtliche Strukturinformationen und Metadaten, die in der XML-Annotation der Ausgangsressource enthalten waren, ist kompatibel mit den im DWDS-Projekt und in DEREKO verwendeten TEI-Schemata und stellt eine Weiterentwicklung der TEI-Modelle aus der TEI-CMCSIG dar.⁴

⁴ Das Schema ist in Form eines ODD-Dokuments im TEI-Wiki dokumentiert und wird als RNG-Datei zur Verfügung gestellt: <http://wiki.tei-c.org/index.php?title=SIG:CMC:clarinschema>. Es fließt in weitere Arbeit der TEI-CMCSIG zur Ausarbeitung eines TEI-Standards für die Repräsentation von Genres internetbasierter Kommunikation ein.

- *Morphosyntaktische Annotation*: Das komplette Korpus wurde mit den Sprachverarbeitungswerkzeugen aus dem Projekt www.schreibgebrauch.de morphosyntaktisch annotiert und auf die Kategorien des Ziel-Tagsets ‚STTS_IBK‘ angepasst. Ein Ausschnitt mit einem Umfang von 4.339 Token wurde von zwei MitarbeiterInnen des Kurationsprojekts manuell nachkorrigiert. Für die manuelle Nachannotation wurde eine von Thomas Schmidt für XML-strukturierte Chat-Daten angepasste Version der Annotationsumgebung *OrthoNormal* (Schmidt 2014) genutzt.
- *Integration der Ressourcen in CLARIN-D*: Die Zielressource wurde in den Repositories an der BBAW und am IDS archiviert⁵ und kann über die Korpusrecherchesysteme des DWDS-Portals⁶ und des DEREKO-Projekts⁷ (COSMAS II bzw. künftig KorAP) vergleichend mit den in den Korpus-sammlungen an der BBAW und am IDS vorhandenen Sprachressourcen (Textkorpora sowie weitere Korpora internetbasierter Kommunikation: Blog-Korpus, Wikipedia-Korpus, Usenet-Korpus) abgefragt und analysiert werden.
- *Rechtsgutachten und Konsequenzen*: Das von *iRights.law* besorgte Rechtsgutachten liefert klare Empfehlungen zur Lizenzierung der Zielressource und zur Überführung der Daten in ein Format, bei dessen Bereitstellung aus rechtlicher Sicht die Risiken so weit reduziert sind, dass die Rechtsabteilungen der beiden CLARIN-D-Zentren bzw. das CLARIN-D LegalHelpDesk die Integration für unbedenklich hielten. Zur Umsetzung der Empfehlungen des Gutachtens war eine aufwändige, großenteils manuell durchgeführte Anonymisierung der Daten erforderlich, die im Anschluss an die eigentliche Projektlaufzeit nach Vorgabe von Harald Längen von studentischen Hilfskräften durchgeführt wurde, die aus Eigenmitteln der Projektleiter finanziert wurden. Die anonymisierte und um ein (kleines) Teilkorpus reduzierte Version der Ziel-Ressource ist ab September 2017 über die o.g. Korpusrecherchesysteme abfragbar.

Die folgenden Abschnitte präsentieren eine Darstellung des Rechtsgutachtens und der daraus gezogenen Konsequenzen. Eine ausführliche Beschreibung der Arbeitsschritte bei der Remodellierung, Annotation und Integration des Korpus in CLARIN-D sowie der dafür entwickelten Lösungen bietet Längen et al. (2016).

⁵ BBAW: <http://hdl.handle.net/11858/00-203Z-0000-002D-EC85-5>,

IDS: <http://hdl.handle.net/10932/00-0379-FDFE-CC30-0301-E>

⁶ <https://www.dwds.de>

⁷ <http://www.ids-mannheim.de/dereko>

3 Rechtliche Bewertung der Ressource mit Blick auf die Integration in CLARIN-D

Gegenstand des Rechtsgutachtens, das im Auftrag des Projekts von der Kanzlei *iRights.law* (John H. Weitzmann und Jan Schallaböck) erstellt wurde, ist die Bewertung der rechtlichen Rahmenbedingungen für die Integration der Ausgangsressource in die CLARIN-D-Korpusinfrastrukturen am Institut für Deutsche Sprache (IDS), Mannheim, und an der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW, Portal www.dwds.de). *iRights.law* ist spezialisiert auf Rechtsfragen im Zusammenhang mit digitalen Medien und dem Internet und kooperiert bereits seit einigen Jahren mit CLARIN-D, beispielsweise bei der Erstellung der DFG-Handreichung zu rechtlichen Aspekten bei der Handhabung von Sprachkorpora (2014).

Die Grundlage für das Rechtsgutachten bildete eine vom Projekt zur Verfügung gestellte Beschreibung der in der Ausgangsressource vorhandenen Teilkorpora mit Erläuterungen zur Herkunft der Daten und zu den Bedingungen ihrer Erhebung und Eingliederung in das Korpus. Darüber hinaus wurden von den Gutachtern stichprobenartig Datenstücke aus den verschiedenen Teilkorpora gesichtet. Den rechtlichen Rahmen für das Gutachten bilden die Regeln zum Datenschutz, zu Persönlichkeitsrechten sowie das Urheberrecht.

Die Gutachter weisen grundsätzlich darauf hin, dass eine verlässliche Prognose der Rechtslage nur bedingt möglich sei, da eine Vielzahl der im Gutachten behandelten Fragestellungen sowohl in datenschutzrechtlicher als auch in urheber- und leistungsschutzrechtlicher Hinsicht bisher nicht höchstrichterlich entschieden seien. Die Empfehlungen des Gutachtens zur Behandlung der Korpusdaten basieren auf einer Abschätzung der rechtlichen Risiken.

3.1 Beschreibung der Ausgangsressource

Die Ausgangsressource für das Projekt bildete das Dortmunder Chat-Korpus, eine Sammlung von Chat-Mitschnitten aus vier verschiedenen sozialen Handlungsbereichen (Freizeit, Bildung, Beratung, Medien) im Umfang von 478 Dokumenten. Jedes Dokument enthält einen Chat-Mitschnitt. Die Gesamtgröße des Korpus beträgt 140.240 Chatter-Beiträge bzw. 1,06 Mio. Tokens. Das Korpus wurde 2002–2008 am Lehrstuhl für Linguistik der deutschen Sprache und Sprachdidaktik der TU Dortmund aufgebaut und ist in der Germanistischen Linguistik und Sprachdidaktik, aber auch im Bereich Computerlinguistik und

Sprachtechnologie als Ressource zur Sprachverwendung in der internetbasierten Kommunikation etabliert.⁸

Das Korpus ist in einer einfachen XML-Sprache annotiert, in der zentrale Strukturmerkmale von Chat-Mitschnitten erfasst und darüber hinaus ausgewählte Sprach- und Stilmerkmale internetbasierter Kommunikation ausgezeichnet sind (unterschiedliche Beitragstypen, darüber hinaus Emoticons, Inflektive, Akronyme, Adressierungen und Nicknames). 383 der insgesamt 478 Dokumente mit einem Umfang von 59.558 Chat-Beiträgen bzw. 548.067 lfd. Wortformen wurden seit unter <http://www.chatkorpus.tu-dortmund.de> sowohl in einer HTML-Version als auch in einer frei downloadbaren XML-Version zur Verfügung gestellt, die neben den Korpusdokumenten ein einfaches, Java-basiertes Abfragewerkzeug (*STACCADo*) umfasste. Die nicht in dieser „Release-Version“ des Korpus enthaltenen Teilkorpora konnten vor Ort in Dortmund genutzt werden.

Das Korpus gliedert sich in verschiedene Teilkorpora, die unterschiedlichen Quellen entstammen und die auf unterschiedliche Weise erhoben wurden. Auf oberster Strukturierungsebene entsprechen die Teilkorpora unterschiedlichen gesellschaftlichen Handlungsbereichen:

- Das Teilkorpus „Chats im E-Learning“ umfasst Mitschnitte aus universitären Lehrveranstaltungen, die entweder komplett virtuell per Chat oder im Blended-Learning-Modus (mit Chat-Anteilen) durchgeführt wurden. Subkorpora repräsentieren Mitschnitte aus unterschiedlichen Lehrveranstaltungen.
- Das Teilkorpus „Chat-Beratung“ umfasst Mitschnitte aus unterschiedlichen Formen chatbasierter Beratungsangebote: von offen zugänglichen, punktuell durchgeführten Großgruppen-Chats einer Studienberatungsstelle über Mitschnitte einer chatbasierten Bibliotheksauskunft, in denen im 1:1-Modus (Ratsuchender und Auskunftsmitarbeiter) Fragen rund um die Bibliotheksrecherche verhandelt werden, bis hin zu 1:1- und Gruppenchats aus Online-Beratungsangeboten zu psychosozialen Themen.
- Das Teilkorpus „Chats im Medienkontext“ umfasst moderierte Chats mit Studiogästen, Politikern und Prominenten, die im Rahmen journalistischer Online-Portale als punktuelle Events durchgeführt wurden. Das Spektrum reicht von Chats mit Politikern aus einem Online-Portal zu politischen Themen über Chats mit Popmusikern, Sportlern und anderen Prominenten in

⁸ Vgl. auch den Beitrag von Dietterle, Lüdeling & Reznicek (in diesem Band).

einem Freizeit-Portal bis hin zu Chats mit Studiogästen im Anschluss an TV-Sendungen.

- Das Teilkorpus „Plauder-Chats“ umfasst zum einen Mitschnitte aus nicht themenbezogenen, unmoderierten Chat-Portalen (Webchats und Internet Relay Chat (IRC)) im Freizeitbereich, die in erster Linie der ungezwungenen Kontakthanbahnung und dem Zeitvertreib genutzt werden. Zum anderen umfasst das Teilkorpus nicht moderierte Chats in journalistischen Online-Portalen („Plauder-Chats im Medienkontext“); dabei handelt es sich i.d.R. um freie Chats im Anschluss an die im Teilkorpus „Chats im Medienkontext“ dokumentierten moderierten Chats, bei denen die Teilnehmer ohne Moderation – im „Plauder-Modus“ – Themen aus den vorangegangenen Events weiterdiskutieren.

Als Grundlage für das Rechtsgutachten wurde eine 20-seitige Dokumentation erstellt, die eine Beschreibung zum Inhalt der einzelnen Teil- und Subkorpora, zum Setting der darin dokumentierten Chats (verwendete Chat-Software, Themenbezug, Moderation etc.) sowie Angaben zu den Bedingungen der Ersterhebung der darin enthaltenen Daten umfassen. Beispiele für drei Korpusteile – ein Subkorporus aus dem Teilkorpus „Chats im E-Learning“, ein Subkorporus aus dem Teilkorpus „Chat-Beratung“, ein Subkorporus aus dem Teilkorpus „Plauder-Chats“ – sind nachfolgend wiedergegeben:

Subkorporus:	1102000
Kurzbeschreibung:	Komplettmitschnitt eines hochschulübergreifenden, virtuellen Seminars zum Thema <i>Lexikographie und Wörterbuchbenutzung</i> (Wintersemester 2002/03)
Setting:	Hochschulübergreifendes Online-Seminar mit Studierenden und Lehrenden dreier deutscher Universitäten. Die Chats waren gerahmt durch eine Reihe an Konversationsregeln, die das Interaktionsmanagement unterstützen und eine kohärente Beitragsabfolge gewährleisten sollten. Diese Regeln wurden den TeilnehmerInnen in der ersten Sitzung vermittelt wurden. Jede Seminarsitzung gliederte sich in eine Großgruppen-Interaktion, eine darauf folgende Phase der Kleingruppenarbeit und abschließend wieder eine Großgruppen-Interaktion. Die Großgruppen-Interaktionen fanden jeweils im Chat-Raum „classroom“ statt; für die Kleingruppenarbeit jedoch begaben sich die Studierenden jeweils in separate Chat-Räume, die die Namen „blue“, „green“ und „red“ trugen. Es gab im Chat einen separaten <i>Teachers</i> -Kanal. Postings, die über diesen Kanal verschickt wurden, wurden regulär in die

Subkorpus:	1102000
	<p>Logfiles am Bildschirm eingeordnet, waren aber nur für Teilnehmer mit Dozenten-Status sichtbar. Sie sind in den Mitschnitten mit dem Zusatz „for teachers“ hinter den Teilnehmernamen gekennzeichnet. Postings, die für alle Beteiligten sichtbar waren, sind mit „for all“ gekennzeichnet.</p> <p>Zum Seminar existierte eine begleitende Website, auf welcher zu einzelnen Sitzungen vorab und/oder nachbereitend Materialien, Begleittexte und Übungsaufgaben bereitgestellt wurden.</p> <p>Verwendetes Chat-System: TULKA</p> <p>Der Seminar-Chatraum war nur für TeilnehmerInnen des Seminars (Lehrende, Studierende) mit gültiger Nutzerkennung zugänglich.</p>
Beispiel-Datensatz:	<p>Großgruppen-Chat (classroom): http://www.chatkorpus.tu-dortmund.de/files/releasehtml/html-korpus/seminar_classroom_30-10-2002_a.html</p> <p>Kleingruppen-Chat (green): http://www.chatkorpus.tu-dortmund.de/files/releasehtml/html-korpus/seminar_green_30-10-2002_a.html</p>
Größe:	Dokumente: 29 Postings: 4.369 Tokens: 28.860
Herkunft der Daten:	Serverseitig erzeugte Mitschnitte sämtlicher Chat-Sitzungen.
Einwilligung:	<p>Die Studierenden und Lehrenden der beteiligten Universitäten gaben zu Semesterbeginn mündlich ihr Einverständnis, dass die Chat-Sitzungen mitgeschnitten und in das in Aufbau befindliche Chat-Korpus integriert werden dürfen. Die Studierenden waren darüber informiert, dass der Aufbau des Korpus für Forschungs- und Lehrzwecke erfolgt und dass die Korpusdaten ForscherInnen sowie der sprachinteressierten Öffentlichkeit online zur Verfügung gestellt werden sollen.</p> <p>Die Mitschnitte wurden am Seminar für Sprachwissenschaft der Universität Tübingen angefertigt, wo auch der Chat-Server betreut wurde. Das Seminar für Sprachwissenschaft stellte die Mitschnitte für die Aufnahme und Bereitstellung im Korpus zur Verfügung.</p>
Subkorpus:	1203000
Kurzbeschreibung:	Chatbasierte Bibliotheksauskunft der Universitätsbibliothek Dortmund
Setting:	Die Universitätsbibliothek Dortmund bietet unter http://www.ub.uni-dortmund.de/literatursuche/auskunftonline.html die Möglichkeit, per Chat 1:1 mit einem Mitarbeiter der Informationsabteilung in

Subkorpus:	1203000
	Kontakt zu treten.
Beispiel-Datensatz:	http://chatkorpus.tu-dortmund.de/files/releasehtml/html-korpus/UB_Gesamt.html
Größe:	Dokumente: 200 Postings: 2.452 Tokens: 21.089
Herkunft der Daten:	Die Mitschnitte wurden von MitarbeiterInnen der UB Dortmund serverseitig erzeugt.
Einwilligung:	Die Mitschnitte wurden vom stellvertretenden Leiter der Universitätsbibliothek für die Aufnahme in das Korpus zur Verfügung gestellt. Die Einwilligung erfolgte mündlich. Die Bibliotheksleitung war darüber informiert, dass der Aufbau des Korpus für Forschungs- und Lehrzwecke erfolgt und dass die Korpusdaten ForscherInnen sowie der sprachinteressierten Öffentlichkeit online zur Verfügung gestellt werden sollen. Ggf. kann eine schriftliche Zustimmung zur Wiederbereitstellung im Rahmen von CLARIN-D nachträglich eingeholt werden.
Anonymisierung:	Die Beteiligten sind in den von der UB zur Verfügung gestellten Ausgangsdaten nicht mit Klarnamen, sondern mit generischen Rollenbezeichnungen identifiziert (<i>AUSKUNFT</i> , <i>BENUTZER</i>).
Subkorpus:	2221000
Kurzbeschreibung:	Freie, unmoderierte Chats in der Webchat-Umgebung „unicum SpaceChat“
Setting:	Freie, unmoderierte Chats in der frei zugänglichen und öffentlich einsehbaren Chat-Umgebung „unicum-SpaceChat“ des Bochumer Uni-Magazins <i>unicum</i> (http://www.unicum.de/). Diese Webchat-Umgebung existiert seit ca. 10 Jahren nicht mehr. Die Chat-Beteiligten konnten ihre Teilnehmernamen (Nicknames) frei wählen.
Beispiel-Datensatz:	http://chatkorpus.tu-dortmund.de/files/releasehtml/html-korpus/unicum_1998.html („Virtuelle Kreuzfahrt“) http://chatkorpus.tu-dortmund.de/files/releasehtml/html-korpus/unicum_21-02-2003_1.html
Größe:	Dokumente: 12 Postings: 15.481 Tokens: 70.825
Herkunft der Daten:	Elf Mitschnitte wurden clientseitig von einer Mitarbeiterin des Korpus-Projekts angefertigt. Ein Mitschnitt („Virtuelle Kreuzfahrt“) wurde von einer Chat-Beteiligten für die Zwecke wissenschaftlicher Untersuchung gespendet.
Einwilligung:	Keine Einwilligung eingeholt; im Falle des gespendeten Chats hat die aufzeichnende Chatterin ihre Einwilligung gegeben, dass der

Subkorpus:	2221000
-------------------	----------------

von ihr zu Erinnerungszwecken mitgeschnittene Chat für Forschungszwecke verwendet werden darf.

3.2 Wesentliche Erkenntnisse und Empfehlungen aus datenschutzrechtlicher Sicht

3.2.1 Personenbezogene Daten

Gegenstandsklärung:

Datenschutzrechtliche Fragestellungen sind grundsätzlich zu beachten, sofern die Korpora personenbezogene Daten enthalten (§ 1 Abs. 1 in Verbindung mit Abs. 2 Bundesdatenschutzgesetz (BDSG)). *Personenbezogene Daten* sind gemäß § 3 Abs. 1 BDSG „Einzelangaben über persönliche oder sachliche Verhältnisse einer bestimmten oder bestimmbaren natürlichen Person (Betroffener)“. Als persönliche Verhältnisse in diesem Sinne sind Angaben über Betroffene selbst, ihre Identifizierung und Charakterisierung anzusehen (z. B. Name, Anschrift, Familienstand, Geburtsdatum, Staatsangehörigkeit, Beruf, Erscheinungsbild, Eigenschaften, Gesundheitszustand). Auch Überzeugungen, Werturteile und biometrische Daten wie Fotografien, Fingerabdrücke sowie die Sprechweise zählen dazu.

Einschätzung der Ressource:

Nach Ansicht der Gutachter ist eine Personenbeziehbarkeit (i.e. Bestimmbarkeit im Sinne der o.g. Norm) nach dem strengen Datenschutzverständnis erst dann nicht mehr gegeben, wenn die Herstellung des Personenbezugs einen klar unverhältnismäßigen Aufwand erfordern würde, von dem nicht zu erwarten ist, dass er getrieben würde.

Soweit die Daten im Korpus auf Veranstaltungen oder Angebote der Auftraggeberinnen selbst zurückgehen (wie z.B. das Subkorpus 1102000 in den Beispiele in Abschnitt 3.1), dürfte es mit einem gewissen Aufwand möglich sein, den einzelnen Chat-Beiträgen die seinerzeit tatsächlich handelnden Personen (z.B. an einem Seminar teilnehmende Studenten) nachträglich zuzuordnen.

Enthalten Chat-Beiträge Einzelangaben über Personen (z.B. über das Alter, die Hobbies, das Aussehen, die Kinder und die täglichen Routinen des Nutzers in Chats), ist von Personenbeziehbarkeit i. S. d. BDSG auszugehen. Auch ein individueller Schreibstil (z.B. die Verwendung von Zahlen anstelle von Buch-

staben, sog. ‚Leetspeak‘) kann prinzipiell mit Kommunikationsbeiträgen außerhalb des Korpus verglichen werden und damit eine Zuordnung zu einer bestimmten Person erlauben. Anders ist dieses Risiko aus stilometrischer Sicht zu bewerten: Aufgrund der Kürze von Chat-Nachrichten dürfte eine mögliche Identifikation über den spezifischen Sprachstil ausgeschlossen sein.

Insgesamt beinhaltet der Datenbestand nach Auffassung der Gutachter wenige systematische Anhaltspunkte für personenbeziehbare Daten. Im Rahmen der Stichprobe konnte kein Fall gefunden werden, in dem es durch Recherchen möglich gewesen wäre, einen Personenbezug herzustellen. Restrisiken bestehen hinsichtlich einer Identifikation im Nahfeld der Person aufgrund unerkannten Zusatzwissens, das eine Zuordnung der Daten ermöglicht.

Empfehlung zur Behandlung der Korpusdaten:

Bei einem Großteil der Korpora wurden bisher keine Standardmaßnahmen (z.B. Randomisierung von Hostnamen; Ersetzung der Nicknames, Ortsnamen, Expertenamen; Anonymisierung, Entfernung oder Permutierung der Zeitstempel) durchgeführt. Dies sollte nach Einschätzung der Gutachter dringend korrigiert werden. Bei Durchführung der Maßnahmen sei zu berücksichtigen, dass für diejenigen, die Zugriff auf die Originalquellen haben (sofern diese noch existieren), die Möglichkeit der Personenbeziehbarkeit fortbesteht. Des Weiteren könne die Durchführung von Standardmaßnahmen zu einer großen Veränderung und in der Folge zu einem Wertverlust der Korpora führen.

Zur Minimierung rechtlicher und ethischer Risiken (insbesondere auch der Gefahr des Reputationsverlustes von CLARIN-D) wird angeraten, ein hohes Datenschutzniveau anzustreben.

3.2.2 Verantwortliche Stelle und anwendbares Recht

Für die Frage, welche Rechtsgrundlagen für die Datenverarbeitung zur Anwendung kommen, ist von zentraler Bedeutung, welche Institution als datenverarbeitende und damit verantwortliche Stelle fungiert. *Datenverarbeitung* ist ein Rechtsbegriff aus § 3 BDSG und erfasst das Speichern, Verändern, Übermitteln, Sperren und Löschen personenbezogener Daten. Sofern das Dachprojekt CLARIN die Verantwortung für die Datenverarbeitung übernimmt, dürften für CLARIN als europäische Forschungsinfrastruktur (ERIC) mit Sitz in Utrecht die nie-

derländischen Datenschutzregelungen anwendbar sein.⁹ Daneben gelten die deutschen Regelungen, sofern man CLARIN-D als Niederlassung von CLARIN im Sinne von Art. 4 Abs. 1 der EU-Datenschutzrichtlinie (95/46/EC) wertet.

Das Hosting des Korpus soll durch das Institut für Deutsche Sprache (IDS) und die Berlin-Brandenburgische Akademie der Wissenschaften (BBAW) durchgeführt werden. Als Stiftung des bürgerlichen Rechts ist das IDS nach Einschätzung der Gutachter trotz teilweiser Beteiligung des Bundes zwar keine öffentliche Stelle im Sinne von § 1 Abs. 2 Nr. 1 BDSG, sondern als nicht-öffentliche Stelle zu qualifizieren; das BDSG bleibt trotzdem anwendbar (vgl. § 1 Abs. 2 Nr. 3 BDSG). Für die BBAW als öffentliche Landeseinrichtung des Landes Berlin¹⁰ gilt hingegen das Berliner Landesdatenschutzgesetz (LDSG Berlin). Auch könnte CLARIN-D als Konsortium als verantwortliche Stelle des Privatrechts agieren; in diesem Fall fänden wiederum die entsprechenden Regelungen des BDSG für nicht-öffentliche Stellen Anwendung.

Für die weitere Beurteilung der rechtlichen Bedingungen unterstellen die Gutachter, dass entweder das IDS oder CLARIN-D (im Folgenden wird vereinfachend nur CLARIN-D genannt) die für die Datenverarbeitung verantwortliche Stelle ist. Das Konsortium hat hierauf durch die Konzipierung der Datenverarbeitung und durch interne Verteilung von Verantwortlichkeiten einen gestaltenden Einfluss, mit der dieses Ergebnis erzielt werden kann.¹¹

Spezialgesetzliche Regelungen (z.B. des Telemediengesetzes) werden als nicht einschlägig erachtet: Zwar erfolgen im Rahmen des Projekts Erhebungen von personenbezogenen Daten im Rahmen eines Telemediendienstes; allerdings beziehen sich diese auf die Nutzer der Plattform, nicht auf die Datenbestände der Korpora, die hier Gegenstand der Untersuchung waren.

9 Beschluss der Kommission vom 29. Februar 2012 über die Einrichtung der gemeinsamen Infrastruktur für Sprachressourcen und -technologien als Konsortium für eine europäische Forschungsinfrastruktur (CLARIN ERIC)

10 Vgl. § 1 Abs. 2 Nr. 2 iVm. Art. 1 Abs. 3 des Staatsvertrages über die Berlin-Brandenburgische Akademie der Wissenschaften.

11 Ein Hosting durch die BBAW oder auch – im Fall CLARIN-D als verantwortliche Stelle – durch das IDS kann dann im Rahmen einer Auftragsdatenverarbeitung nach § 11 BDSG durchgeführt werden, für die dort weitere Zulässigkeitsvoraussetzungen normiert sind, die aber für die Frage der grundsätzlichen Rechtmäßigkeit der Erhebung und Verarbeitung hier außer Betracht bleiben können.

3.2.3 Rechtmäßigkeit der Datenverarbeitung

Das BDSG verlangt prinzipiell für jeder Verarbeitung personenbezogener Daten eine Rechtsgrundlage, sofern für die konkrete Datenverarbeitung eine solche nicht besteht, ist die Verarbeitung rechtswidrig.¹² In Betracht käme eine für die untersuchten Fälle zum einen die Einwilligung nach § 4a BDSG oder der Erlaubnistatbestand des § 29 BDSG.¹³ Diese beiden Rechtsgrundlagen werden nachfolgend näher ausgeführt.

3.2.3.1 Einwilligung nach § 4a BDSG

Sofern und soweit ein Eingriff in die datenschutzrechtlich gesicherten Bereiche nicht durch Anonymisierung ausgeschlossen werden kann, braucht es eine Rechtsgrundlage. Eine solche könnte in einer Einwilligung bestehen.

- a) *Mehrere Beteiligte:* Bei bi- und multilateraler Kommunikation, die in geschlossenen Benutzergruppen stattgefunden hat, gilt zunächst zu beachten, dass die Einwilligung aller Beteiligter vorliegen muss. Außerdem ist zu berücksichtigen, dass eine Einwilligung der unmittelbar Teilnehmenden die Daten, die diese Chattenden über Dritte verbreiten, regelmäßig nicht umfasst. Die Herstellung eines Personenbezugs bei diesen Daten würde allerdings einen äußerst hohen Aufwand erfordern, sofern Stichproben diesbezüglich repräsentativ waren. Diesbezüglich wird also unterstellt, dass keine Personenbeziehbarkeit im datenschutzrechtlichen Sinne vorliegt.
- b) *Schriftlichkeit:* Gemäß § 4a Abs. 1 Satz 3 BDSG ist grundsätzlich für jeden weiteren Nutzungsschritt in Bezug auf die Daten eine schriftliche Einwilligung (eigenhändig unterzeichnet, klar umgrenzt) erforderlich, also insbesondere auch zur Veröffentlichung in einem anderen Kontext.
- c) In Ausnahmefällen kann eine andere Form angemessen sein (vgl. § 4a Abs. 1 Satz 3 BDSG). Grundsätzlich wäre nach Auffassung der Gutachter eine

¹² Manchmal in begrifflich fragwürdiger Weise auch in der datenschutzrechtlichen Fachliteratur als „Verbot mit Erlaubnisvorbehalt“ beschrieben. Tatsächlich handelt es sich um eine klassische rechtliche Regel-Ausnahme-Technik, wie sie in vielen Rechtsbereichen vorzufinden ist. Der Begriff des „Verbots mit Erlaubnisvorbehalt“ ist ein Terminus technicus des Verwaltungsrechts und unterscheidet verschiedene Formen von Genehmigungstatbeständen. Er dient der Abgrenzung des präventiven Verbots mit Erlaubnisvorbehalt vom repressiven Verbot mit Befreiungsvorbehalt..

¹³ Entsprechend der oben dargestellten Empfehlung, die Verarbeitung in Verantwortung durch das IDS oder CLARIN-D selbst durchzuführen, konzentriert sich die Prüfung auf die Tatbestände der §§ 27ff. BDSG und lässt solche für den öffentlichen Bereich außer Betracht.

konkludente Einwilligung in Form schlüssigen Verhaltens denkbar, da Teilnehmer in Chats, Newsgroups, Foren und anderen netzbasierten Kommunikationsformen üblicherweise deren Funktionsweise und die damit verbundene öffentliche Sichtbarkeit der Beiträge kennen; eine Verwendung in anderem Kontext wäre mangels Vorhersehbarkeit allerdings auch von einer konkludenten Einwilligung nicht erfasst.

- d) *Minderjährigenschutz*: Für eine wirksame Einwilligung Minderjähriger ist deren jeweilige Einsichtsfähigkeit abhängig von Alter, geistiger Reife und Verwendungszusammenhang maßgeblich. Eine Orientierung bieten die Altersgrenzen der §§ 2, 104 ff. BGB.
- e) Da Alter und geistige Reife der im Korpus erfassten Chatter mitunter nicht zweifelsfrei feststellbar sind, verbleibt auch bei vorliegenden Einwilligungen ein erhebliches Restrisiko. Dieses erhöht sich zusätzlich, wenn davon ausgegangen wird, dass neben einer wirksamen Einwilligung eines Minderjährigen auch die Einwilligung aller Erziehungsberechtigten vorliegen muss (sog. „Theorie von der Doppelzuständigkeit“).
- f) *Freie Widerruflichkeit*: Es gilt zu beachten, dass datenschutzrechtliche Einwilligungen für die Zukunft frei widerruflich sind, was in Landesdatenschutzgesetzen teils ausdrücklich geregelt ist, nach herrschender Meinung aber verfassungsrechtlich begründet auch für das BDSG gilt¹⁴, in dem die Widerruflichkeit etwa in Hinblick auf spezielle Verwendungsprivilegien in § 28 und darüber hinaus mit weiteren Anforderungen in § 35 Abs. 5 ausdrücklich genannt wird.
- g) *Datenübermittlung im Ausland*: Zum Zeitpunkt der Begutachtung konnte nicht beurteilt werden, inwieweit in Zukunft aus dem Ausland auf die Korpora zugegriffen werden soll. In einem solchen Fall wären die besonderen Zustimmungs- und Einwilligungsregelungen des § 4b BDSG zu beachten.

Fazit zur Einwilligung nach § 4a BDSG: Eine allein auf Einwilligungen der Betroffenen gestützte Datenverarbeitung ist nach Ansicht der Gutachter aufgrund erheblicher rechtlicher und tatsächlicher Unsicherheiten nicht empfehlenswert.

3.2.3.2 Erlaubnistatbestand des § 29 BDSG

Ein *Erlaubnistatbestand* ist eine gesetzliche Regelung, die ein an sich verbotenes Handeln im Einzelfall ausnahmsweise gestattet. Da CLARIN-D die Korpora

¹⁴ Vgl. T. Herbst, MedR (2009) 27, S. 149 (150) m. w. N.

Dritten zur Verfügung stellen will, ist Gegenstand der Prüfung nicht die eigene Forschungstätigkeit von CLARIN-D¹⁵, sondern die Frage der Zulässigkeit und des Umfangs der Übermittlung der Korpora. Dass CLARIN-D als Forschungsinfrastruktur einen unterstützenden Beitrag zur Forschung anderer leisten will, ändert nichts daran, dass der Schwerpunkt im Übermittlungszweck liegt.¹⁶

Entscheidend ist damit, ob die Voraussetzungen des § 29 BDSG (Geschäftsmäßige Datenerhebung und -speicherung zum Zwecke der Übermittlung) erfüllt sind.

Für die Tätigkeit von CLARIN-D ist von *Geschäftsmäßigkeit* auszugehen. Geschäftsmäßigkeit ist dadurch definiert, dass die Tätigkeit auf eine gewisse Dauer angelegt und auf Wiederholung gerichtet ist, was bei CLARIN-D der Fall ist. Für die weitere Beurteilung der Rechtmäßigkeit sind verschiedene Phasen der Verarbeitung zu unterscheiden: Die erste Erhebung durch eine andere Stelle, der Datenzufluss, die Verarbeitung im engeren Sinne, sowie die anschließende Übermittlung an Dritte.¹⁷

- a) *Ersterhebung*: Für eine Rechtswidrigkeit der Ersterhebungen der Daten beim Betroffenen bestehen nach Auffassung der Gutachter keine Anhaltspunkte.
- b) *Datenzufluss*: Datenzufluss bezeichnet den Vorgang, während dessen die Daten nach ihrer Erhebung zum Datenverarbeiter (hier CLARIN-D) gelangen. Voraussetzung für die Rechtmäßigkeit der Erhebung durch CLARIN-D ist zunächst, dass kein schutzwürdiges Interesse der Betroffenen an einem Ausschluss der Datenverarbeitung besteht, § 29 Abs. 1 Nr. 1 BDSG, oder dass sie aus einer „allgemein zugänglichen Quelle“ stammen und entgegen stehende berechnete Interessen der Betroffenen nicht *offensichtlich* sind, § 29 Abs. 1 Nr. 2 BDSG.

Für problematisch hielten die Gutachter diesbezüglich das Teilkorpus 1201000, das Mitschnitte eines psychosozialen Beratungsangebots enthält, in denen Ratsuchende u.a. in Einzelchats psychische Belastungen, Störungen und Suchterkrankungen mit Beratern thematisieren. Entsprechend der

¹⁵ In einem solchen Fall müsste zunächst eine Prüfung des § 28 BDSG erfolgen.

¹⁶ § 28 BDSG (Datenerhebung und -speicherung für eigene Geschäftszwecke) scheidet aus diesem Grund als Rechtsgrundlage aus. Ähnliches gilt hinsichtlich der Verweisung des § 29 Abs. 5 BDSG auf § 28 Abs. 6 BDSG, dort insbesondere Nr. 4. Hiernach wäre nur eine Verarbeitung zu einem konkreten Forschungsvorhaben gedeckt, was nach angezeigter enger Auslegung nicht für ein Infrastrukturprojekt wie CLARIN-D gelten kann. Das Gleiche gilt für die Sondervorschrift des § 40 BDSG (Verarbeitung und Nutzung personenbezogener Daten für Forschungszwecke) zur Veröffentlichung von Daten.

¹⁷ Ehmann in Simitis, Kommentar zum BDSG, 8. Auflage 2014, § 29. Rn. 121.

Sensitivität dieser Gegenstände bestehe ein hohes schutzwürdiges Interesse der Betroffenen, weshalb anzunehmen ist, dass bereits die Erfassung des Korpus durch CLARIN-D nicht datenschutzrechtskonform wäre. Zwar mag aus der reinen Erhebung durch CLARIN-D noch kein massiver Eingriff erwachsen, wohl aber wäre dies – mangels zugesicherter Anonymisierung – im Falle der Weitergabe der Bestände gegeben. Sofern aber eine Weitergabe unzulässig ist, entfällt auch der Zweck der Erhebung, was auch diese unzulässig macht. Allenfalls kann die Erhebung noch auf den Zweck gestützt werden, dass man eine Eignung für die spätere Übermittlung prüfen wolle. Spätestens mit der abschlägigen Prüfung entfällt aber die Erforderlichkeit der weiteren Speicherung. Aus diesem Grund empfehlen die Gutachter dringend, das Teilkorpus 1201000 aus der in CLARIN-D zu integrierenden Version des Korpus zu entfernen sowie alle Kopien zu löschen.

Für alle anderen Teilkorpora ist für die Gutachter (nach Stichprobenprüfung) ein überwiegendes Interesse der Betroffenen, das einer Erhebung durch CLARIN-D entgegenstünde, nicht erkennbar.

- c) *Verarbeitung im engeren Sinne*: Die im Projekt vorgesehenen Aufbereitungen der Korpora (Konvertierungen, Einbringen von Annotationen etc.) stellen keine Datenverarbeitungen dar, die die geschützten Interessen der Betroffenen in *besonderer* Weise tangieren. Derartige Verarbeitungen treffen auf keine datenschutzrechtlichen Bedenken.
- d) *Übermittlung an Dritte*: Für die Integration der Korpora kommt es damit darauf an, ob die Übermittlungszwecke gegenüber den berechtigten Interessen der (potentiell) Betroffenen überwiegen bzw. offensichtlich überwiegen. Im Rahmen der CLARIN-D-Forschungsinfrastruktur werden Datenbestände in verschiedenen Modellen angeboten, die nach dem Grad der Zugänglichkeit der ursprünglichen Datenquellen zu unterscheiden sind:
 - *Korpora aus allgemein verfügbaren Quellen*:¹⁸ Für diese ist im Idealfall eine öffentliche Verfügbarmachung angestrebt (CLARIN-Endnutzer-Lizenz vom Typ PUB), die u.a. keinen spezifischen datenschutzrechtli-

¹⁸ Teilweise umstritten ist dabei die hier durchaus relevante Frage, ob auch eine Kommunikationsumgebung, die lediglich einer Anmeldung bedarf, die aber jedermann zugänglich ist, noch als allgemein verfügbare Quelle zu qualifizieren ist. Schon zur Vermeidung von rechtlichen Risiken, sei hier empfohlen, solche Inhalte nicht als Inhalte aus allgemein verfügbaren Quellen zu betrachten. Sachlich spricht hierfür im Übrigen auch, dass diese Inhalte regelmäßig nicht durch Suchmaschinen indiziert werden können und sich damit tatsächlich deutlich von den üblichen öffentlich zugänglichen Inhalten unterscheiden.

chen Beschränkungen unterliegt.¹⁹ Im Falle einer Übermittlung unter dieser Lizenz würde es sich um eine Übermittlung zum Zwecke der allgemeinen Zugänglichmachung handeln. Rechtlich würde CLARIN-D hier eine einer Suchmaschine ähnliche Funktion übernehmen. Auf Basis der Stichprobenprüfung gehen die Gutachter davon aus, dass hinsichtlich derjenigen Korpora, die aus allgemein zugänglichen Quellen stammen, die geschützten Interessen der Betroffenen jedenfalls nicht offensichtlich überwiegen. Voraussetzung hierfür wäre allerdings die Durchführung von Standardmaßnahmen (vgl. 3.2.1). Unter dieser Bedingung können diese Korpora öffentlich zugänglich gemacht werden.

- *Korpora aus geschlossenen Nutzergruppen:* Ausweislich § 29 Abs. 1 Nr. 1 und Abs. 2 Nr. 2 BDSG darf bei dieser Art der Korpora kein Grund zur Annahme bestehen, dass der Betroffene ein schutzwürdiges Interesse am Ausschluss der Verarbeitung und Übermittlung hat. Ein solches könnte sich daraus ergeben, dass sich der Nutzer ganz bewusst nicht frei für jedermann einsehbar, sondern in einer geschlossenen Nutzergruppe geäußert hat. Dabei ist auch zu berücksichtigen, dass bei einer öffentlichen Zugänglichmachung nicht auszuschließen ist, dass die Datenbestände durch Dritte weiter indiziert werden und nach einiger Zeit auch über Suchmaschinen gefunden werden können. So wäre schließlich denkbar, dass eine in einem geschlossenen Chat gemachte Äußerung – möglicherweise noch dazu weitgehend zusammenhangslos – in einer Suchanfrage zu einer Person (einem gewählten Pseudonym, einer sprachlichen Eigenheit) auftaucht. Eine freie Veröffentlichung dieser Teilkorpora außerhalb des wissenschaftlichen Kontexts scheint daher datenschutzrechtlich nicht angeraten.

Sofern die Verfügbarmachung des Korpus auf Forschungszwecke eingeschränkt wird, stehen einer Übermittlung keine schutzwürdigen Interessen der Betroffenen entgegen, weshalb eine Zugänglichmachung datenschutzrechtlich möglich erscheint.

- *Abwägung:* Die frühere Rechtsprechung des BGH verlangte eine Abwägung zwischen den schutzwürdigen Belangen des Persönlichkeitsrechts des Betroffenen gegen die Interessen der speichernden Stelle. Nach Änderung des Wortlauts des BDSG verlangen einige weiterhin

¹⁹ Vgl. dazu <https://corpora.uni-hamburg.de/drupal/de/korpusanfragen-lizenzen>, allerdings gelten naturgemäß auch für diese Datenbestände und die Datenverarbeitung durch die Empfänger der Bestände datenschutzrechtliche Anforderungen, sofern sie diesen gesetzlich unterworfen sind.

diese am Verhältnismäßigkeitsgrundsatz ausgerichtete Abwägung.²⁰ An den vorstehenden Ergebnissen ändert sich nach Auffassung der Gutachter allerdings auch dann nichts, wenn man dem § 29 Abs. 1 Nr. 1 und Abs. 2 Nr. 1 BDSG ein Abwägungselement hinzufügen wollte. Anders als für bereits allgemein zugängliche Daten könne gerade an dem in geschlossenen Nutzergruppen formulierten Beitrag kein pauschales allgemeines Informationsinteresse aus Art. 5 Abs. 1 Grundgesetz (GG) hergeleitet werden, welches das Recht am eigenen Wort überwiegen könnte. Die auch die Informationsfreiheit umfassende Meinungsfreiheit schützt ausweislich ihres Wortlautes nur die Information aus allgemein zugänglichen, nicht die aus geschlossenen Quellen. Allerdings könne bei einer Abwägung zwischen dem Recht am eigenen Wort und der Forschungsfreiheit aus Art. 5 Abs. 3 GG die Letztere als überwiegend unterstellt werden, so dass auch nach dieser Auffassung das gleiche Ergebnis erzielt wird.

Des Weiteren verweisen die Gutachter auf § 28 Abs. 1 Satz 2 BDSG als eigene Zulässigkeitsvoraussetzung. Demnach seien bei der Erhebung der Daten die konkreten Zwecke der Verarbeitung festzulegen. Dabei reiche eine reine Festlegung auf den „Zweck der Übermittlung“ nicht aus. Einige Stimmen in der Literatur fordern für die Privilegierung zu Forschungszwecken gemäß § 28 Abs. 3 Nr. 4 BDSG eine konkrete Bestimmung des Forschungsvorhabens²¹, andere sehen bei klarer Definition von Ziel, Aufbau und Verlauf des Vorhabens auch das Anlegen von langfristigen Forschungsdatenbanken zur Entwicklung wissenschaftlicher Infrastruktur²² erfasst. Sofern man – der Rechtsprechung folgend – für § 29 Abs. 1 Nr. 2 BDSG eine Zweckkonkretisierung im Sinne des Ansinnens, „der Öffentlichkeit die Nutzung des Internets zu erleichtern, wenn nicht gar erst zu ermöglichen“ als hinreichende Anforderung gelten lassen wolle, müsse das erst recht gelten, wenn Wissenschaftlern die Nutzung von Forschungsdaten erleichtert oder gar erst ermöglicht werden solle. Dies ergibt sich nach Ansicht der Gutachter systematisch nicht zuletzt daraus, dass Forschungszwecke im Datenschutzrecht regelmäßig Privilegierungen genießen, aber auch – jedenfalls wertend – aus der grundrechtlichen Forschungsfreiheit des Art. 5 Abs. 3 GG. Somit spräche vieles dafür, dass § 29 BDSG nicht deshalb als mögliche Rechtsgrundlage für die Verarbeitung der Schriftkorpora durch CLARIN-D auszu-

²⁰ Vgl. zum Streitstand: Ehmann in Simitis, § 29 Rn. 453ff.

²¹ AaO, § 28 Rn. 203.

²² Dammann in Simitis, § 14 Rn. 91.

schließen sei, weil die konkreten Forschungszwecke nicht hinreichend bestimmt seien. Allerdings wird ausdrücklich darauf hingewiesen, dass andere Auffassungen zu dieser Problematik vertretbar sind, weshalb eine Verfügbarmachung zu Forschungszwecken zwar vertretbar erscheint, aber nicht den Weg des geringsten rechtlichen Risikos darstellt.

Empfehlung zur Behandlung der Korpusdaten:

Im Rahmen von CLARIN-D kann nach Einschätzung der Gutachter das Risiko reduziert werden, indem

- 1) die öffentliche Verbreitung auf diejenigen Teilkorpora beschränkt wird, die unter § 29 Abs. 1 Nr. 2 BDSG fallen, d. h. die insbesondere Daten aus allgemein zugänglichen Quellen enthalten;
- 2) alle Elemente, die nicht nur unerhebliche Identifizierungsrisiken bergen, im Rahmen der bereits empfohlenen Anonymisierung unkenntlich gemacht werden. Als mögliche Verfahren werden empfohlen: Variablenunterdrückung, Objektunterdrückung, Informationsreduktion, Zufallsvertauschungen und Zufallsüberlagerungen. Eine für jeden Einzelfall verlässliche automatische Anonymisierung von Freitexten erscheint dagegen technisch nicht möglich, da diese mit maschinellen Lernverfahren arbeiten, die bei der Übertragung der Muster auf andere Chats versagen.

Die Anwendung der Rechtsgrundlage des § 29 BDSG erfordere weiterhin:

- eine Glaubhaftmachung des berechtigten Interesses, § 29 Abs. 2 Nr. 1 BDSG;
- eine Aufzeichnung der Übermittlungen, § 29 Abs. 4 BDSG;
- eine Meldung der Verarbeitung bei der zuständigen Aufsichtsbehörde, § 4d Abs. 1 BDSG.

Bei denjenigen Teilkorpora, die im Sinne von § 19 UrhG öffentlich zugänglich gemacht werden, sei laut Gutachten das berechtigte Interesse als allgemeines Informationsinteresse stets indiziert, eine Überprüfung und eine Protokollierung der Zugriffe könne in diesen Fällen entfallen. Aus Gründen der Datensparsamkeit erscheine es sogar geboten, auf eine solche Protokollierung zu verzichten.²³

²³ Das Prinzip der Datensparsamkeit besagt, dass von einem Betroffenen nur insoweit Daten erhoben und verarbeitet werden dürfen, als unbedingt notwendig.

3.3 Wesentliche Erkenntnisse und Empfehlungen aus persönlichkeitsrechtlicher Sicht

Das Allgemeine Persönlichkeitsrecht (APR; Art. 2 Abs. 1 in Verbindung mit Art. 1 Abs. 1 GG) jedes Menschen ist insbesondere ein Verfügungsrecht über die Darstellung der Person in der Öffentlichkeit und bildet die Basis für den Schutz der Sozialsphäre, Privatsphäre und Intimsphäre.²⁴

Soweit allerdings Datenschutznormen eingreifen (siehe oben), gehen diese als speziellere Regelungen der unmittelbaren Anwendung des APR vor.

Das APR ist folglich nur als Auffangregelung anzusehen. Ist es anwendbar, muss eine umfassende Interessenabwägung zwischen der Intensität des Eingriffs in den Schutzbereich des APR und dem öffentlichen Interesse an der Zugänglichmachung der Information (Art. 5 GG) vorgenommen werden: So sei nach Auffassung der Gutachter bei Beiträgen in öffentlichen Foren eine relativ geringe Beeinträchtigung anzunehmen, insoweit sich die Äußernden freiwillig ihres Persönlichkeitsschutzes begeben hätten.

Eine besondere Ausprägung des APR ist der Schutz der persönlichen Ehre: Diesbezüglich merken die Gutachter an, dass allein auf Beleidigung und Herabwürdigung erkennbarer Dritter abzielende Beiträge aus den Korpora gefiltert werden müssten, da ansonsten eine Verbreiterhaftung drohe. Das bedeutet, dass jeder Rückbezug auf eine bestimmbar Person ausgeschlossen werden muss, z.B. durch eine Anonymisierung.

Eine weitere Ausprägung des APR, die insbesondere bei Äußerungen in geschlossenen Nutzergruppen Bedeutung gewinnt, sei das Recht am eigenen Wort: Geschützt ist hierdurch die Selbstbestimmung über die eigene Darstellung der Person in der Kommunikation mit anderen; auch die Befugnis zur Bestimmung des Empfängerkreises des Kommunikationsinhalts ist erfasst. Jedoch trete der Schutz des Rechts am eigenen Wort auch subsidiär hinter die Datenschutzregelungen zurück, weshalb für eine Anwendung lediglich der Bereich der *allgemeinen Rede* verbleibe.²⁵ Ein Eingriff in diesen Schutzbereich läge allerdings nur dann vor, wenn diese Rede realistisch einer bestimmten wirklichen Person zugeordnet werden könne.

²⁴ Vgl. die infolge der Rechtsprechung des BVerfG entwickelte Sphärentheorie (Sozialsphäre, Privatsphäre und Intimsphäre).

²⁵ Der Bereich der *allgemeinen Rede* umfasst im juristischen Sinne alles Gesagte bzw. Geschriebene mit Ausnahme von Angaben im Sinne von Informationen, Tatsachenbehauptungen oder Werturteilen.

Mangels Bild-, Video- und Tondateien in den Korpora finden §§ 22, 23 des Gesetzes betreffend das Urheberrecht an Werken der bildenden Künste und der Photographie (KUG) und das – aus dem APR abgeleitete – Recht an der eigenen Stimme vorliegend keine Anwendung. Auch eine Verletzung des Namensrechts aus § 12 BGB scheidet nach Ansicht der Gutachter mangels Namensbestreitung oder Namensanmaßung aus, da die Verwendung von Nicknames gerade keinen Anspruch des Verwenders impliziert, tatsächlich so zu heißen.

3.4 Wesentliche Erkenntnisse und Empfehlungen aus urheber- bzw. leistungsschutzrechtlicher Sicht

3.4.1 Urheberrecht: Chat-Beiträge als Sprachwerke

Nach dem Dafürhalten der Gutachter sind allenfalls einige wenige Textbeiträge innerhalb der Teilkorpora der Ausgangsressource gemäß § 2 Abs. 1 Urheberrechtsgesetz (UrhG) urheberrechtlich schutzfähig. Trotz der geringen Anforderungen gemäß dem Maßstab der sog. „kleinen Münze“²⁶ als persönlicher geistiger Schöpfung im Sinne des § 2 Abs. 2 UrhG scheiden die meisten Chat-Beiträge aufgrund ihrer Kürze und/oder eines Mangels an Individualität aus.

Sollte ausnahmsweise Urheberrechtsschutz bestehen, würden durch das Hosting der Korpora v.a. das Vervielfältigungsrecht (§ 16 UrhG), das Recht der öffentlichen Zugänglichmachung (§ 19a UrhG) und das Veröffentlichungsrecht (§ 12 UrhG) tangiert. Folglich bedürfte es dann eines wirksamen Nutzungsrechts oder einer eingreifenden *Urheberrechtsschranke*²⁷. Sofern Urheberrechtsschutz an den Beiträgen besteht (siehe oben), wäre die Speicherung auf den CLARIN-D-Servern als eine Vervielfältigung (§ 16 UrhG), die nur mit Erlaubnis des Urhebers zulässig ist. Die Schrankenregelung des § 44a UrhG zugunsten technisch bedingter Zwischenspeicherungen greife vorliegend nicht ein.

- a) Im Fall geschlossener Interaktionen (private Chatrooms) könne mit der Einbindung in die CLARIN-D-Forschungsstruktur eine erstmalige öffentliche Wiedergabe im Sinne des § 15 Abs. 2 UrhG vorliegen. Öffentlichkeit liegt

²⁶ Um Urheberrechtsschutz genießen zu dürfen, muss es ein Werk nicht zwangsläufig herausragend sein. Auch durchschnittliche Werke sind dem Schutz zugänglich, sofern sie die erforderliche Individualität aufweisen.

²⁷ Grundsätzlich darf ein urheberrechtlich geschütztes Werk nur mit Zustimmung des Urhebers genutzt werden. Urheberrechtsschranken erlauben hingegen unter besonderen Umständen von Gesetzes wegen eine Nutzung des Werkes, um die Interessen der Allgemeinheit bzw. bestimmter Nutzergruppen (z. B. der Forschung) in angemessenen Ausgleich zu bringen.

gemäß § 15 Abs. 3 UrhG immer dann vor, wenn diejenigen Personen, denen ein Werk wahrnehmbar gemacht wird, nicht untereinander oder mit demjenigen, der die Wiedergabe vornimmt, „persönlich verbunden“ sind. Die über CLARIN-D auf Korpora zugreifenden Personen seien – auch im Falle strikter Zugangskontrollen – keineswegs in der vorgenannten Weise miteinander verbunden.

Die Textinhalte aus geschlossenen, etwa bilateralen, Chats würden somit durch ein Einstellen bei CLARIN-D im rechtlichen Sinne erstmals öffentlich wiedergegeben, was – soweit urheberrechtlicher Schutz besteht (siehe oben) – neben dem Nutzungsrecht nach § 19a UrhG das Veröffentlichungsrecht des Urhebers aus § 12 Abs. 1 UrhG berührte. Dies gälte unabhängig davon, ob die jeweilige Urheberin bzw. der Urheber dank Anonymisierung nicht mit wirklichem Namen, etwa nur unter einem Pseudonym, erkennbar ist. Eine ohne Zustimmung erfolgte Entfernung des Namens könne sogar eine weitere, vertiefende Verletzung des Anerkennungsrechts des Urhebers aus § 13 UrhG darstellen.

- b) Bei zahlreichen Teilkorpora des Dortmunder Chat-Korpus liegen bzw. lagen lediglich mündliche Erlaubnisse vor. Im Streitfalle gäbe es zunächst ein erhebliches Nachweisproblem (es wäre richterliche Überzeugung vom Vorliegen der Erlaubnis notwendig, § 286 Zivilprozessordnung (ZPO)). Inwieweit das in concreto der Fall wäre, entzieht sich der Bewertung der Gutachter, da sie die beteiligten Personen und genauen Umstände nicht näher kennen. Fraglich sei ferner, wie weit die mündliche Rechteeinräumung reiche: Jedenfalls sei ein einfaches Nutzungsrecht zur Vervielfältigung (§ 16 UrhG) sowie zur öffentlichen Zugänglichmachung (§ 19a UrhG) eingeräumt worden. Dieses sei nicht auf ein bestimmtes Forschungsvorhaben und damit auch zeitlich und örtlich unbeschränkt. Allerdings beschränke sich der Nutzungszweck auf Sprachforschung und allgemeine Sprachrecherche. Eine Unterlizenzierung sei nach Einschätzung der Gutachter nicht erlaubt.
- c) Die in der CLARIN-D-Infrastruktur gebotenen Möglichkeiten, Annotationen vorzunehmen und die Inhalte zu analysieren, verändern nach Ansicht der Gutachter die eigentlichen Interaktionen inhaltlich nicht. Daher liege insofern eine freie Benutzung im Sinne von § 24 Abs. 1 UrhG vor, die unabhängig von etwaigen Erlaubnissen urheberrechtlich zulässig sei.
- d) Bezüglich der Korpora, bei denen keine Erlaubnisse und damit keine Nutzungsrechteinräumungen der Beitragenden vorliegen, wäre ein Rückgriff auf die Schranken des Urheberrechts zu erwägen. Ein öffentliche Zugänglichmachung unter der Schranke des § 52a UrhG wäre lediglich denkbar, wenn die Zugriffsmöglichkeit auf einen bestimmt abgegrenzten Kreis von

Personen für deren eigene wissenschaftliche Forschung beschränkt würde, was allerdings erhebliche Auswirkungen auf das technische Design der CLARIN-D-Infrastruktur hätte. Die Schranke des § 53 Abs. 2 Nr. 1 UrhG wiederum sei auf die Sammlungen einzelner Wissenschaftler ausgerichtet und daher für den vorliegenden Fall ungeeignet.

Empfehlung zur Behandlung der Korpusdaten:

Angesichts des geringen Anteils potenziell schutzfähiger Inhalte, der ebenfalls geringen Wahrscheinlichkeit, dass deren Rechteinhaber wirklich materiell in Urheberinteressen berührt sind und ein rechtliches Vorgehen²⁸ auch realistisch erwägen würden, und der selbst dann eher geringen wirtschaftlichen Risiken, stellt sich nach Einschätzung der Gutachter der Umgang mit dem Urheberrecht der Chat-Teilnehmenden eher als eine Frage des öffentlichen Ansehens und des Vorbildcharakters der beteiligten Institutionen dar denn als eine juristische.

Wenn sicher ausgeschlossen werden sollte, dass eine urheberrechtsverletzende Integration in die CLARIN-D-Infrastruktur erfolgt, müssten die Textinhalte vollständig – Zeile für Zeile – fachlich begutachtet werden –, und selbst dann könnten Anteile übrig bleiben, bei denen erneut nur eine Risikoabschätzung zu der Frage, wie ein Gericht den Sachverhalt im Streitfall sehen könnte, möglich sei.

Folgende Vorsichtsmaßnahmen könnten demgegenüber ein rechtliches Risiko zumindest zusätzlich verringern: Für längere (nicht dialogische) Passagen könnte die Möglichkeit des Downloads unterbunden werden oder bei Abfrage des Systems nur eine ausschnittsweise Wiedergabe der gefundenen Stellen erfolgen. Denkbar wäre auch ein reines Query-and-analysis-only (QAO), bei dem gar kein Einblick in die Daten gewährt wird; dies hätte allerdings höchstwahrscheinlich eine gravierende Einschränkung der Forschungsarbeit zur Folge.

28 In Betracht käme ein strafrechtliches (§§ 106 ff. UrhG) sowie ein zivilrechtliches Vorgehen (§§ 97–105 UrhG). Selbst auf eine Strafanzeige hin wäre es extrem unwahrscheinlich, dass sich eine Staatsanwaltschaft in Deutschland dazu veranlasst sehen könnte, Forschungsinstitutionen wegen unberechtigter Zugänglichmachung von Forschungsmaterial zu verfolgen. Zivilrechtlich wäre eine Abmahnung der beteiligten Institutionen denkbar: Aber selbst bei einem hoch angesetzten Gegenstandswert von 10.000 Euro längen die Gebühren gemäß Rechtsanwaltsvergütungsgesetz (RVG, 1,0-Gebühr nach Nr. 2300 des Vergütungsverzeichnisses) bei etwas über 500 €. Der zu addierende (verschuldensabhängige) Schadensersatz würde nominell ausfallen, da es keinen Lizenzmarkt für Chat-Äußerungen gibt, und wäre praktisch vernachlässigenswert.

3.4.2 Datenbankherstellerrecht (§§ 87ff. UrhG)

Gemäß § 87b UrhG ist es ohne Zustimmung des Herstellers einer Datenbank grundsätzlich unzulässig, einen nach Art und Umfang wesentlichen Teil einer Datenbank oder die Datenbank als Ganze zu vervielfältigen, zu verbreiten oder öffentlich wiederzugeben. Grundvoraussetzung für einen solchen Anspruch wäre das Vorliegen einer Datenbank im Rechtssinne. Erforderlich ist hierfür im Sinne von § 87a UrhG eine Datensammlung, deren Elemente systematisch und methodisch angeordnet sind und mit Hilfe elektronischer Mittel einzeln zugänglich gemacht werden und für deren Beschaffung, Überprüfung und Darstellung eine wesentliche Investition erforderlich war bzw. ist. Ob Chats oder andere dialogische Kommunikationsumgebungen Datenbanken in diesem Sinne darstellen, wurde bisher höchstrichterlich noch nicht entschieden. Davon sei nach Ansicht der Gutachter aufgrund der einfachen Funktionsweise der automatischen, chronologischen Anordnung der Konversationen mangels Leistung des Datenbankherstellers nicht auszugehen. Selbst wenn eine thematische Strukturierung existiere, erfolge diese durch die Chat-Teilnehmenden, nicht durch den Systembetreiber als Datenbankhersteller.

Eine wesentliche Investition im Sinne von §§ 87a ff. UrhG läge hingegen in der Entwicklung der für die Chats genutzten Software. Deren Rechteinhaber hätten nach Einschätzung der Gutachter jedoch voraussichtlich kein Interesse an einer Verfolgung einer Verletzung ihres Datenbankherstellerrechts – falls ein solches überhaupt bestehe.

Anders zu bewerten sei die Problematik hingegen mit Blick auf die eigenen Datenbankherstellerrechte der Institutionen, die die Aufbereitung der Korpora initiiert und durchgeführt haben. Bezüglich des Dortmunder Chat-Korpus wäre dies vermutlich die TU Dortmund. Empfehlenswert sei hierfür – sofern die Korpora freigegeben werden sollen – eine Creative-Commons-Lizenz der Version 4.0, die das Leistungsschutzrecht des Datenbankherstellers mitlizenziert, so dass bei Nachnutzung die bekannten CC-Lizenzbedingungen wie Namensnennung, ShareAlike usw. gelten.²⁹

²⁹ Bei Lizenzversion 3.0. wird das Datenbankherstellerrecht durch Freigabe mittels der Lizenzen aufgegeben.

3.4.3 Erwägungen zur Lizenzwahl

Die bei CLARIN in den drei Zugriffskategorien PUB (öffentlich verfügbar), ACA (Nutzer muss einer akademischen Einrichtung angehören) und RES (Beschränkte Nutzung; zusätzliche Anforderungen müssen erfüllt sein) eingeordneten Lizenzen und Erklärungen sind durchweg Standardvereinbarungen, die vor allem Urheber- und Leistungsschutzrechte lizenzieren.³⁰

Sofern überhaupt Rechte an Teilen der Korpora bestehen, scheide nach Empfehlung der Gutachter der Einsatz einer Aufgabenerklärung (Creative Commons-Lizenz CC0: Urheber überlässt Werk der Public Domain für freien Gebrauch) aus, da die auftraggebenden Institutionen – soweit ersichtlich – an keinem der betroffenen Inhalte vollständig über alle Rechte verfügen könnten.

Für die Korpora, an denen Datenbankherstellerrechte der involvierten Institutionen bestehen, stünden grundsätzlich alle Lizenzoptionen unter CLARIN-PUB zur Wahl. Bei der Wahl der konkreten Lizenz sei allerdings zu beachten, dass die Creative-Commons-Lizenzen bis einschließlich Version 3.0 das Datenbankherstellerrecht entweder ganz aussparen (gilt für die Fassungen „Generic“ bzw. „Unported“) oder sogar einen ausdrücklichen Verzicht auf dieses Recht enthalten (gilt für mehrere der für europäische Rechtsordnungen portierten Fassungen). Ab Version 4.0 sei das anders; eine Entscheidung für eine Creative-Commons-Lizenz der Version 4.0 erleichtere überdies die Nachnutzbarkeit der Ressource.

4 Konsequenzen für die Aufbereitung der Korpusdaten

Gemäß der Empfehlung d.) in 3.2.3.2 wurde das Teilkorpus 1201000 (psychosozialer Beratungsschat) vollständig aus dem Bestand gelöscht. Das Korpus wird im Rahmen der CLARIN-D-Forschungsinfrastrukturen (Repositorien und Korpus-researchsysteme an den Institutionen BBAW und IDS) zur Verfügung gestellt. Um die weiteren Empfehlungen zu befolgen, wird das Teilkorpus mit denjenigen Chats, die aus geschlossenen Chat-Plattformen stammen, ausschließlich registrierten und akademisch affilierten Nutzern zur Verfügung gestellt.

³⁰ Vgl. Oksanen et al. (2010). Das APR und das gesamte Datenschutzrecht werden von diesen Lizenzen nicht geregelt.

4.1 Anonymisierung

Sowohl zur Minimierung rechtlicher und ethischer Risiken im Datenschutz (Abschnitt 3.2.1) als auch um zu untermauern, dass der Erlaubnistatbestand des §23 BDSG für die Verarbeitung des Chatkorpus angenommen werden kann (Abschnitt 3.2.3.2), empfiehlt das Gutachten Standardmaßnahmen zur Anonymisierung personenbezogener Daten in den Chat-Protokollen bzw. eine Unkenntlichmachung aller Elemente, die nicht nur unerhebliche Identifizierungsrisiken bergen. Beispiele für personenbezogene Daten werden in 3.2.1 genannt.

Es wird argumentiert, dass eine Personenbeziehbarkeit dann nicht mehr gegeben ist, wenn ihre Herstellung einen klar unverhältnismäßigen Aufwand erfordern würde. Um also den erforderlichen Aufwand so weit wie möglich, d.h. unter Wahrung der Brauchbarkeit des Korpus für linguistische Untersuchungen, zu erhöhen, werden Standardmaßnahmen zur Anonymisierung empfohlen, als Beispiele werden genannt: Randomisierung von Hostnamen; Ersetzung der Nicknames, Ortsnamen, Expertennamen; Anonymisierung, Entfernung oder Permutierung der Zeitstempel.

Weiterhin werden in 3.2.3.2 Variablenunterdrückung, Objektunterdrückung, Informationsreduktion, Zufallsvertauschungen und Zufallsüberlagerungen als mögliche Anonymisierungsmethoden genannt. Diese Verfahren werden in Höhne (2010) genauer definiert, sie beziehen sich allerdings auf personenbezogene, u.a. numerische Mikrodaten, die in einer oder in mehreren aggregierbaren Datenbank-Tabellen vorliegen. Im Folgenden orientieren wir uns vorwiegend an der Literatur zur Anonymisierung von textuellen Daten. Dort spricht z.B. Medlock (2006) von *Kategorisierung* (was einer Variablenunterdrückung mit Ersatzinformation nach Höhne (2010) zu entsprechen scheint) und von *Pseudonymisierung* (ein Vertauschungs- bzw. Austauschungsverfahren).

Um einen Überblick a.) über weitere mögliche Typen von sensiblen Referenzen, d.h. referenziellen sprachlichen Ausdrücken im Chatkorpus, die potenziell anonymisiert werden müssen und b.) über mögliche Verfahren zur Unkenntlichmachung oder Neutralisierung der sensiblen Referenzen zu gewinnen, sichteten wir aktuelle Literatur zu Anonymisierungsverfahren für Sprachkorpora sowie Beschreibungen von durchgeführten Anonymisierungen in Projekten zum Aufbau von IBK-Korpora, darunter zwar keine Chatkorpora, aber ein E-Mail-Korpus (Medlock 2006), ein Facebook-Korpus (Didi 2015), zwei SMS-Korpora (Pankhurst et al. 2013, Ueberwasser 2015), sowie eines Gesprächskorpus (Winterscheid 2015).

4.1.1 Kategorien

Damit ein Korpus noch auf vernünftige Weise für linguistische Untersuchungen verwendet werden kann, empfiehlt Medlock (2006), dass sensible Referenzen nicht einfach entfernt werden, sondern *kategorisiert*, d.h. ersetzt werden durch einen Platzhalter, der die Kategorie des ersetzten Elements (wie Personenname oder Ortsname) wiedergibt, oder sogar, *pseudonymisiert*, d.h. ersetzt werden durch eine Variante aus derselben Kategorie (beispielsweise indem der Vorname *Holger* durch *Werner* ersetzt wird). Anonymisierung ist dann der Oberbegriff für Kategorisierung, Pseudonymisierung oder weitere mögliche Verfahren. Der Aufwand für eine ordentliche Pseudonymisierung von sensiblen Referenzen ist allerdings deutlich höher als der für eine Kategorisierung, da eine Konsistenz der Ersetzungen auch bzgl. Formvarianten gewährleistet sein sollte (*Werners* für *Holgers*, *weeerners* für *hooolgers* etc.), wofür ausführliche Maskierungstabellen (vgl. Winterscheid 2015) erstellt werden müssen. Im Falle des Chatkorpus wurde daher, wie in den von uns betrachteten anderen IBK-Korpusprojekten, die Anonymisierung als Kategorisierung durchgeführt; dies schließt natürlich eine spätere Pseudonymisierung, um die Nutzbarkeit noch weiter zu verbessern, nicht aus.

Da die meisten Elemente, die im Chatkorpus anonymisiert werden mussten, Namen sind, berufen wir uns zunächst auf die fünf *Named Entity*-Kategorien PER, ORG, LOC, GPE, OTH, die in der TüBa-D/Z Baumbank (Telljohann et al. 2004) annotiert wurden (vgl. Tabelle 1) und die bereits in Experimenten zur *Named Entity Recognition* für DeReKo verwendet wurden (Bingel/Haider 2014). Da die fünf Kategorien dieses Sets recht weit gefasst sind und in einigen Fällen die vorhandenen Annotationen der Ausgangsressource bereits spezifischere Informationen enthielt, erweiterten wir das Set um die Kategorien NICK (für Nickname, als Unterkategorie von PER) und ROOM (für den Namen eines Chatrooms). Zusätzlich wurde die Kategorie GEO_DE für Substantive oder Adjektive, die von einem Namen (zumeist LOC oder GPE) abgeleitet sind (*Dortmunder*) (GEO_DE entspricht einer Vereinigung der beiden Kategorien _GeoNE_ und _GeoADJA_ aus DiDi 2015).

Neben den Kategorien für Namen werden drei weitere Kategorien für formellere Referenzen eingeführt: URL (für eine WWW-Adresse), EMAIL (für eine E-Mail-Adresse) und NUMBER (für jegliche Art von referenzierender Nummer, Beispiele s. Tabelle 1). Angelehnt an Winterscheid (2015) bzw. DiDi (2015) führen wir die beiden selteneren, aber dennoch relevanten Kategorien IMPLICIT (für eine implizite Referenz) und CITATION (für ein Zitat, das zur Identifizierung eines Individuums verwendet werden könnte) ein.

Alle 13 verwendeten Anonymisierungskategorien sind in Tabelle 1 mit genaueren Definitionen, Beispielen und Angabe der Quelle bzw. Anregung für die Kategorie aufgeführt.

Tab. 1: Anonymisierungskategorien im Dortmunder Chatkorpus 2.0.

#	Kurzform	Langform und Definition/Beispiele Beschreibung/ Erläute- rung	Quelle bzw. Anregung für Kategorie	#Vorkommen nach Anony- misierung, Σ = 39519
NAMEN				
1	PER	PERSONNAME Personenname: Vorna- me oder Nachname oder Sequenz aus Vorname und Nachna- me	TüBaDZ (Telljohann et al. 2004)	838
2	NICK	NICKNAME Nickname: Benutzer- name, den ein Chatteil- nehmer sich gegeben hat, oder eine Variante davon). NICK ist eine Unterkategorie von PER, also immer wenn ein Nickname erkannt wurde, wurde NICK annotiert, ansonsten PER.	Dortmunder Chat Korpus	30022
3	ORG	ORGANISATIONNAME Organisationsname: Firma (z.B. Arbeitgeber eines Teilnehmers), Sportverein, Institut, Universität etc. NB: Bei einem Ausdruck wie “Universität Hil- desheim” wurde nur “Hildesheim” anonymi-	TüBaDZ	741

#	Kurzform	Langform und Definition/Beispiele Beschreibung/ Erläute- rung	Quelle bzw. Anregung für Kategorie	#Vorkommen nach Anony- misierung, $\Sigma = 39519$
		siert (Universität ist ein hinreichend allgemei- ner Ausdruck), und dieses dann als GPE.		
4	GPE	GEOPOLITICALENTITY- NAME Geo-Political Entity: Ort oder Gebiet, dessen Grenzen politisch- verwaltungsmäßig festgelegt sind, also alle Städte, Gemein- den, Länder, Bundes- länder, Stadtteile etc. inkl. Schreibvarianten und Abkürzungen davon.	“Mannheim”, “NRW”, “Italien” “doaaadmund” “DO” TüBaDZ	1309
5	GEO_DE	GEODERIVATIONNAME Geo-Derivation: Sub- stantiv oder Adjektiv, welches von einem Namen, meist GPE oder LOC, morphologisch abgeleitet ist und die Zugehörigkeit oder Eigenschaft (Adjektiv) oder eine Gruppe oder Bewohner (Substantiv) bezeichnet.	“Mannheimer”, “Mannheimerin- nen”, “Gelbfüßler” “Schwabe” “sommer- lath'sche” “doadmunder” DiDi (2015)	178
6	LOC	LOCATIONNAME Location: Ort oder Gegend, der oder die nicht politisch definiert ist, z.B. Berg, Tal, Fluss, Straße, Autobahn etc.	“Augustaanlage”, “Neckar”, “Königstuhl” “A6” TüBaDZ	107
7	OTH	OTHERNAME Other: Restkategorie für alles was sonst	“bravo” (=Name einer TüBaDZ	1819

#	Kurzform	Langform und Definition/Beispiele Beschreibung/ Erläuterung	Quelle bzw. Anregung für Kategorie	#Vorkommen nach Anonymisierung, $\Sigma = 39519$
		nicht zugeordnet werden kann (Namen und Referenzen, z.B. auch Bars)	Zeitschrift bzw. einer Chat-Plattform)	
8	ROOM	ROOMNAME Name eines Chatraums	“Welcome”, “emerald” DO Chat- korpus	2409

ÜBRIGE REFERENZEN

9	URL	WWWURL WWW-Adresse	“http://www.ids-mannheim.de/”	1742
10	EMAIL	EMAIL E-Mail-Adresse	“sommer-lath@ids-mannheim.de”	50
11	NUMBER	NUMBER Nummer: Jegliche Zahl oder Nummer oder Kodierung, die auf eine Person oder einen Gegenstand, der einer Person zugeordnet werden kann verweist: Hausnummer, Seriennummer, Postleitzahl, Telefonnummer, Passnummer, Kontonummer, IP-Adresse, Passwort.	“0621/1581418”, “10.0.1.81”, “68161”	169
12	IMPLICIT	IMPLICIT Implizite Referenz: Aussagekräftige Beschreibungen und Informationen, aus denen Rückschlüsse auf die Identität eines Chat-Teilnehmers oder eines Dritten gezogen werden kann. Inklusive	Sein Hobby ist das Sammeln von „Weinflaschenetiketten“, „IT-Operator“	130

#	Kurzform	Langform und Definition/Beispiele Beschreibung/ Erläute- rung	Quelle bzw. Anregung für Kategorie	#Vorkommen nach Anony- misierung, $\Sigma = 39519$
		Berufsbezeichnungen, falls sie einen Hinweis auf eine Person geben.		
13	CITATION	CITATION Zitat: Zitate einer ande- ren Person, oder aus einem Werk, das zur Identifizierung eines Teilnehmers oder eines Dritten verwendet werden könnte.		5

4.1.2 Technische Umsetzung

Die Ausgangsressource enthielt in den Metadaten jedes Chatprotokolls bereits eine Liste der Chatteilnehmer („creatorList“) mit Angabe einer ID, ihres Geschlechts und ihrer Rolle im Chat (wie „teacher“ oder „expert“). In den eigentlichen Chat-Texten waren Nicknames und Anreden anderer Teilnehmer (@Willi) bereits annotiert, in der Regel mit einer Verlinkung in die Teilnehmerliste. Ebenso waren die Namen der Chaträume als solche annotiert (sie erscheinen vorwiegend in Statusmeldungen des Chat-Systems). Diese vorhandenen Annotationen wurden in das TEI-Format des Kurationsprojekts (‘CLARIN D-TEI-CMC’, vgl. Abschnitt 2) konvertiert und bildeten den Ausgangspunkt für Schritt 1 der Anonymisierung. Die Teilnehmerliste erscheint im CLARIN-D-TEI als <particList>, im eigentlichen Chat-Textbereich sind Nicknames mit <name type=“NICK“> annotiert, Adressierungen mit <ref type=“addressingTerm“> und die Referenzierung in die <particList> des TEI-Headers erfolgt TEI-konform per @who bzw. @corresp.

Die Dateinamen enthielten teilweise die Namen von Chatplattformen; sie wurden über eine Mapping-Liste anonymisiert, die die Teilstrings mit sensiblen Referenzen auf thematische Bezeichner wie „Politik-Portal“ abbildete. Die Chat-Teilnahme oder Erwähnung von Personen des öffentlichen Lebens wie Politikern und Prominenten gilt als nicht-sensibel, die Vorkommen solcher Namen wurden also nicht anonymisiert.

Schritt 1 - Automatische Anonymisierung durch ein XSLT-Stylesheet: Der String-Inhalt aller bereits annotierten Nicknames und Adressierungen wurden getilgt und durch einen kategorisierenden Platzhalter ersetzt. Ein Platzhalter-String bestehen aus folgenden drei Teilstrings: a.) dem Inhalt des @sex-Attributes des Chat-Teilnehmer, aus dem entsprechenden <person>-Element in der <particList> im TEI-Header.; b.) dem Inhalt des @role-Attributes des Teilnehmers, oder falls es kein @role gab, der String „PARTICIPANT“ und c.) der @xml:id des Teilnehmers (vgl. Listing 1). Diese „sprechenden“ Ersetzungs-Strings wie „FEMALE-TEACHER-A08“ ersetzen den textuellen Inhalt der <name>-Annotationen wie auch der Attribute @lemma und @normal (normalisierte Form) eines zu anonymisierenden <w>-Elements, sowie auch den Textinhalt von <persName> im TEI-Header in der Teilnehmerliste selbst (Listing 1). Sollte die Annotation von <name> oder <ref> keine Referenzierung in die <particList> enthalten, so können ID, Geschlecht und Rolle nicht automatisch ermittelt werden und der Ersetzungsstring lautet lediglich „NICKNAME“. Beispiele für das Zielformat nach dem automatischen Anonymisierungsschritt finden sich in Listing 1, 2 und 4.

Schritt 2 - Manuelle Anonymisierung: In einer Kampagne mit vier Hilfskräften der Universität Mannheim, der Universität Duisburg-Essen und des IDS wurde in dem gesamten Chatkorpus verbleibende sensible Referenzen wie unter Abschnitt 4.1.1 definiert mit Hilfe des „Author Mode“ des XML-Editors Oxygen identifiziert und mit dem Kategorienlabel-Inventar aus Tabelle 1 annotiert. Das Resultat wird derzeit geprüft und dann durch ein weiteres Stylesheet in die TEI-Annotation analog zu Schritt 1 überführt: Namen werden mit <name>, übrige Referenzen mit <ref> umgesetzt, dabei wird die Kurzform des Kategorienlabels im @type-Attribut festgehalten und die Langform im Ersetzungs-String; der Ersetzungs-String erscheint wiederum als Textinhalt und als Inhalt der Attribute @lemma und @normal, wenn vorhanden. Beispiele für die Umsetzung von manuell annotierten Kategorien finden sich in den Listings 3 und 4.

Das Chatkorpus 2.0 enthält rund 1,06 Millionen Tokens, davon wurden nun insgesamt 39519 Einheiten (Wortbestandteile, Wörter oder Mehrworteinheiten) anonymisiert. Wie oft eine Anonymisierungskategorie zugewiesen wurde, zeigt die letzte Spalte von Tabelle 1.

Listing 1: Anonymisierung von Metadaten (Teilnehmerliste). Personen des öffentlichen Lebens (Prominente, Politiker) werden nicht anonymisiert.

```

<particDesc>
  <!-- 1301005 -->
  <listPerson>
    <!-- ... -->
    <person role="celebrity" xml:id="A03">
      <persName type="NICK">Günther Beckstein</persName>
      <sex evidence="estimated">male</sex>
    </person>
    <!-- ... -->
    <person role="moderator" xml:id="A04">
      <persName type="NICK">[_ MALE-MODERATOR-A04_]</persName>
      <sex evidence="estimated">male</sex>
    </person>
    <!-- ... -->
    <person role="participant" xml:id="A07">
      <persName type="NICK">[_ FEMALE-PARTICIPANT-A07_]</persName>
      <sex evidence="estimated">female</sex>
    </person>
    <person role="participant" xml:id="A08">
      <persName type="NICK">[_ PARTICIPANT-A08_]</persName>
      <sex evidence="estimated">unknown</sex>
    </person>
    <!-- ... -->
  </listPerson>
</particDesc>

```

Listing 2: Anonymisierung eines Nicknames ohne Rolleneintrag in Teilnehmerliste

```

<post auto="false" rend="color:lime" type="event" who="#A14" xml:id="m487">
  <name corresp="#A14" type="NICK">
    <w lemma="[_ PARTICIPANT-A14_]" type="NE" xml:id="m487.t1">
[_ PARTICIPANT-A14_]</w>
  </name>
  <w lemma="werden" type="VAFIN" xml:id="m487.t2">wird</w>
  <w lemma="schlecht" type="ADJD" xml:id="m487.t3">schlecht</w>
</post>

```

Listing 3: Anonymisierung einer Ableitung von einem Namen (wie *Düsseldorfern*)

```

<w lemma="auch" type="ADV" xml:id="m40.t1">auch</w>
<w lemma="bei" type="APPR" xml:id="m40.t2">bei</w>
<w lemma="die" type="ART" xml:id="m40.t3">den</w>
<name type="GEO_DE">
  <w lemma="[_GEODERIVATIONNAME_]" type="NN"
xml:id="m40.t4">[_GEODERIVATIONNAME-4_]</w>
</name>

```

Listing 4: Anonymisierung einer impliziten Referenz (Studienfach eines Teilnehmers)

```

<post auto="false" rend="color:#808080" synch="#t427" type="standard"
who="#A26" xml:id="m576">
  <w lemma="wollen" type="VMFIN" xml:id="m576.t1">willst</w>
  <w lemma="du" type="PPER" xml:id="m576.t2">du</w>
  <w lemma="eine" type="ART" xml:id="m576.t3">ne</w>
  <w lemma="Therapie" type="NN" xml:id="m576.t4">therapie</w>
  <ref corresp="#A31" type="addressingTerm">
    <w lemma="@" type="ADRIND" xml:id="m576.t5">@</w>
    <w type="NE" xml:id="m576.t6">[_MALE-PARTICIPANT-A31_]</w>
  </ref>
  <w lemma="ich" type="PPER" xml:id="m576.t7">ich</w>
  <w lemma="studieren" type="VVFIN" xml:id="m576.t8">studier</w>
  <ref type="IMPLICIT">
    <w type="NE" xml:id="m576.t9">[_IMPLICIT-1_]</w>
  </ref>
</post>

```

4.2 Lizenzwahl

Da die Datenbankherstellerrechte (3.4.2) für die Aufbereitung zumindest zu einem Teil bei dem CLARIN-D-Kurationsprojekt liegen, und da die Verantwortlichen für den Aufbau der ursprünglichen Ressource gleichzeitig Mitglieder des Kurationsprojekts waren, wurde die Ressource gemäß der Empfehlung in 3.4.3 mit dem Lizenztyp CC-BY 4.0 versehen.

5 Fazit und Ausblick

Die unklare rechtliche Situation in Bezug auf die Speicherung, Aufbereitung und Wiederbereitstellung von Sprachdaten internetbasierter Kommunikation ist einer der wesentlichen Gründe dafür, weshalb bislang kaum frei verfügbare Korpora zu diesem wichtigen Kommunikationsbereich existieren. Für die Einschätzung der rechtlichen Situation und der daraus für die Auswahl und Repräsentation von Korpusdaten erwachsenden Konsequenzen fehlt Linguisten, Korpus- und Computerlinguisten die juristische Expertise.

In diesem Beitrag berichteten wir über die Ergebnisse eines Rechtsgutachtens, das im Rahmen des CLARIN-D-Kurationsprojekts *ChatCorpus2CLARIN* in Auftrag gegeben wurde und das die Bedingungen für die Integration eines Korpus mit bereits erhobenen Chat-Daten in die Korpusinfrastrukturen des BMBF-Verbundprojekts CLARIN-D klären sollte.

Die Erstellung des Rechtsgutachtens gestaltete sich als sehr aufwändig. Zwar wurde das Gutachten durch einen externen Auftragnehmer – die Kanzlei *iRights.law* – erarbeitet; auch verfügte der Auftragnehmer bereits über einschlägige Expertise in Bezug auf digitale Daten. Um eine geeignete Basis für die Einschätzung der Ressource zu schaffen, war allerdings eine Erweiterung der vorhandenen Dokumentation zu den Korpusdaten und waren insbesondere mehrere Rückfrage- und -klärungsschleifen zwischen Projektbeteiligten und Juristen erforderlich, so dass die endgültige Fassung des Gutachtens erst kurz vor Projektende vorlag.

Die Ergebnisse des Gutachtens machten es erforderlich, einen Teil des Datenbestands aus der Ressource zu entfernen. Die übrigen Korpusdaten wurden von Grund auf neu anonymisiert. Für die Anonymisierung wurden Best Practices aus existierenden Textkorpora sowie aus anderen Korpora internetbasierter Kommunikation aufgegriffen und für Zwecke des Chat-Korpus angepasst.

Ein schon länger existierendes Korpus nachträglich rechtlich zu prüfen, wie im hier berichteten Projekt der Fall, ist sicherlich um einiges aufwändiger als die rechtliche Situation vor Beginn der Datensammlung einzuschätzen und davon ausgehend erforderliche Weichenstellungen bereits auf der Ebene der Erhebung von Daten und Metadaten vorzunehmen. Dennoch ist auch bei Korpusprojekten, die erst am Anfang stehen, davon auszugehen, dass, wie auch im vorliegenden Fall, für eine Einschätzung der rechtlichen Situation am konkreten Einzelfall verschiedene rechtliche Teilgebiete gegeneinander abgewogen werden müssen (Datenschutzrecht, Persönlichkeitsrecht, Urheber- und Leistungsschutzrecht).

Hilfreich für die Konzeption von Korpora bzw. von jeder Art datengestützter Forschung auf Basis von Daten internetbasierter Kommunikation bzw. aus ‚sozialen Medien‘ wäre die Entwicklung von rechtlichen Leitlinien, die beim Umgang mit solchen Daten grundsätzlich zu bedenken sind. Die Entwicklung solcher Leitlinien kann nicht im Rahmen eines einzelnen Korpusprojekts geleistet werden und bedarf der Mitwirkung von Juristen. Sie könnte eine lohnenswerte Aufgabe für Ressourceninfrastrukturinitiativen wie CLARIN und ihre nationalen Partnerprojekte sein.

Literatur

- Bartz, Thomas, Michael Beißwenger und Angelika Storrer (2014): 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* 28 (1), 157–198. Online unter: http://www.jlcl.org/2013_Heft1/7Bartz.pdf (21.07.2017).
- Beißwenger, Michael (2013): Das Dortmunder Chat-Korpus. In: *Zeitschrift für germanistische Linguistik* 41/1, 161–164. Erweiterte Fassung auf dem Linguistik-Server Essen (LINSE): Online unter: <http://tinyurl.com/chatkorpus> (21.07.2017).
- Beißwenger, Michael, Thomas Bartz, Angelika Storrer und Swantje Westpfahl (2015): Tagset und Richtlinie für das Part-of-Speech-Tagging von Sprachdaten aus Genres internetbasierter Kommunikation. Guideline-Dokument aus dem Projekt "GSCL Shared Task: Automatic Linguistic Annotation of Computer-Mediated Communication / Social Media" (EmpiriST2015). Online unter: <http://https://sites.google.com/site/empirist2015/home/annotation-guidelines> (21.07.2017).
- Beißwenger, Michael, Sabine Bartsch, Stefan Evert und Kay-Michael Würzner (2016): EmpiriST 2015: A Shared Task on the Automatic Linguistic Annotation of Computer-Mediated Communication and Web Corpora. In: *Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task*. Stroudsburg: Association for Computational Linguistics (ACL Anthology W16-2606), 44–56. Online unter: <http://aclweb.org/anthology/W/W16/W16-2606.pdf> (21.07.2017).
- Beißwenger, Michael, Maria Ermakova, Alexander Geyken, Lothar Lemnitzer und Angelika Storrer (2012): A TEI Schema for the Representation of Computer-mediated Communication. *Journal of the Text Encoding Initiative*, (3) (doi: 10.4000/jtei.476). Online unter: <http://jtei.revues.org/476> (21.07.2017).
- Bingel, Joachim und Thomas Haider (2014): Named Entity Tagging a Very Large Unbalanced Corpus. Training and Evaluating NE Classifiers. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC '14)*. Reykjavik: European Language Resources Association (ELRA)
- Chanier, Thierry, Céline Poudat, Benoît Sagot, Georges Antoniadis, Ciara Wigham, Linda Hriba, Julien Longhi und Djamé Seddah (2014): The CoMeRe corpus for French: structuring and annotating heterogeneous CMC genres. *Journal of language Technology and Computa-*

- tional Linguistics* 29 (2), 1–30. Online unter: http://www.jlcl.org/2014_Heft2/1Chanier-et-al.pdf (21.07.2017).
- DFG (2014): Informationen zu rechtlichen Aspekten bei der Handhabung von Sprachkorpora. Bonn: Deutsche Forschungsgemeinschaft.
- DiDi (2015): Beschreibung der Anonymisierung im DiDi-Korpus. Online unter: http://www.eurac.edu/en/research/autonomies/commul/Documents/DiDi/DiDi_anonymisation_DE.pdf (21.07.2017).
- Fišer, Darja und Michael Beißwenger (Eds.) (2016): Proceedings of the 4th Conference on CMC and Social Media Corpora for the Humanities (cmc-corpora2016). Ljubljana, Slovenia, 27–28 September 2016. Online unter: <http://nl.ijs.si/janes/cmc-corpora2016/proceedings/> (21.07.2017).
- Herbst, Tobias (2009): Die Widerruflichkeit der Einwilligung in die Datenverarbeitung bei medizinischer Forschung: MedR – Medizinrecht, 2009, 149–151 (doi: 10.1007/s00350-009-2361-6).
- Höhne, Jörg (2010): Verfahren zur Anonymisierung von Einzeldaten. Reihe Statistik und Wissenschaft, Band 15. Hrsg. Statistisches Bundesamt Wiesbaden.
- Horbach, Andrea, Diana Steffen, Stefan Thater und Manfred Pinkal (2014): Improving the Performance of Standard Part-of-Speech Taggers for Computer-Mediated Communication. Proceedings of KONVENS 2014, 171–177.
- Lüngen, Harald, Michael Beißwenger, Axel Herold und Angelika Storrer (2016): Integrating corpora of computer-mediated communication in CLARIN-D: Results from the curation project ChatCorpus2CLARIN. In: Dipper, Stefanie, Friedrich Neubarth und Heike Zinsmeister (eds.): Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016), 156–164. Online unter: https://www.linguistics.rub.de/konvens16/pub/20_konvensproc.pdf (21.07.2017).
- Lüngen, Harald und C.M. Sperberg-McQueen (2012): A TEI P5 Document Grammar for the IDS Text Model. In: Journal of the Text Encoding Initiative (JTEI), Issue 3 | November 2012. Online unter: <http://jtei.revues.org/508> (DOI: 10.4000/jtei.508 (21.07.2017).
- Margaretha, Eliza und Harald Lüngen (2014): Building Linguistic Corpora from Wikipedia Articles and Discussions. In: Journal of language Technology and Computational Linguistics (JLCL) 29 (2), 59–82. Online unter: http://www.jlcl.org/2014_Heft2/3MargarethaLuengen.pdf (21.07.2017).
- Medlock, Ben (2006): An introduction to nlp-based textual anonymization. In: Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006), 1051–1056.
- Oksanen, Ville, Krister Lindén und Hanna Westerlund (2010): Laundry symbols and license management: Practical considerations for the distribution of LRs based on experiences from CLARIN. In: Proceedings of LREC 2010: Workshop on Language Resources: From Storyboard to Sustainability and LR Lifecycle Management. Malta.
- Panckhurst, Rachel (2013): A large SMS Corpus in French: from Design and Collation to Anonymisation, Transcoding and Analysis. In: Procedia - Social and Behavioral Sciences 95, 96–104.
- Schmidt, Thomas (2014): The Research and Teaching Corpus of Spoken German – FOLK. In: Proceedings of the Ninth conference on International Language Resources and Evaluation (LREC'14). Reykjavic, Iceland: European Language Resources Association (ELRA). Online unter: http://www.lrec-conf.org/proceedings/lrec2014/pdf/290_Paper.pdf (21.07.2017).

- Simitis, Spiros (2014) Bundesdatenschutzgesetz. 8. Aufl.. Wiesbaden: Nomos, zit: *Bearbeiter* in Simitis.
- Telljohann, Heike, Erhard Hinrichs und Sandra Kübler (2004): The TüBa-D/Z Treebank: Annotating German with a Contet-Free Backbone. In: Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004).
- Ueberwasser, Simone (2015): Anonymisation (SMS4Science.ch). Online unter: <https://sms.linguistik.uzh.ch/bin/view/SMS4Science/Anonymisation> (21.07.2017).
- Westpfahl, Swantje (2014): STTS 2.0? Improving the Tagset for the Part-of-Speech-Tagging of German Spoken Data. In: Lori Levin & Manfred Stede (Hrsg.): Proceedings of LAW VIII – The 8th Linguistic Annotation Workshop. Dublin, Ireland: Association for Computational Linguistics and Dublin City University, 1–10. Online unter: <http://www.aclweb.org/anthology/W14-4901> (21.07.2017).
- Westpfahl, Swantje und Thomas Schmidt (2016): FOLK-Gold – A GOLD standard for Part-of-Speech-Tagging of Spoken German. In: Proceedings of the Tenth conference on International Language Resources and Evaluation (LREC'16)
- Winterscheid, Jenny (2015): GAIS Web: Maskierung. Dokument im IDS-Prowiki Online unter: <http://prowiki.ids-mannheim.de/bin/view/GAIS/MasKierung> (21.07.2017).
- Zinsmeister, Heike, Ulrich Heid, und Kathrin Beck (Hrsg.) (2014): Das STTS-Tagset für Wortarttagging – Stand und Perspektiven. Themenheft, Journal for Language Technology and Computational Linguistics 28 (1).