Aufbau und Annotation eines Referenzkorpus zur deutsch-sprachigen internetbasierten Kommunikation (*DeRiK*-Projekt)

technische universität dortmund

Michael Beißwenger

berlin-brandenburgische AKADEMIE DER WISSENSCHAFTEN

Lothar Lemnitzer

Workshop: Webkorpora in Computerlinguistik und Sprachforschung // Mannheim, 27./28. Sept. 2012

GSCL

INSTITUT FÜR DEUTSCHE SPRACHE

Ziel: Aufbau eines Referenzkorpus zur deutschsprachigen internetbasierten Kommunikation:

„Deutsches Referenzkorpus zur internetbasierten Kommunikation“ (DeRiK)

Kooperationsprojekt:

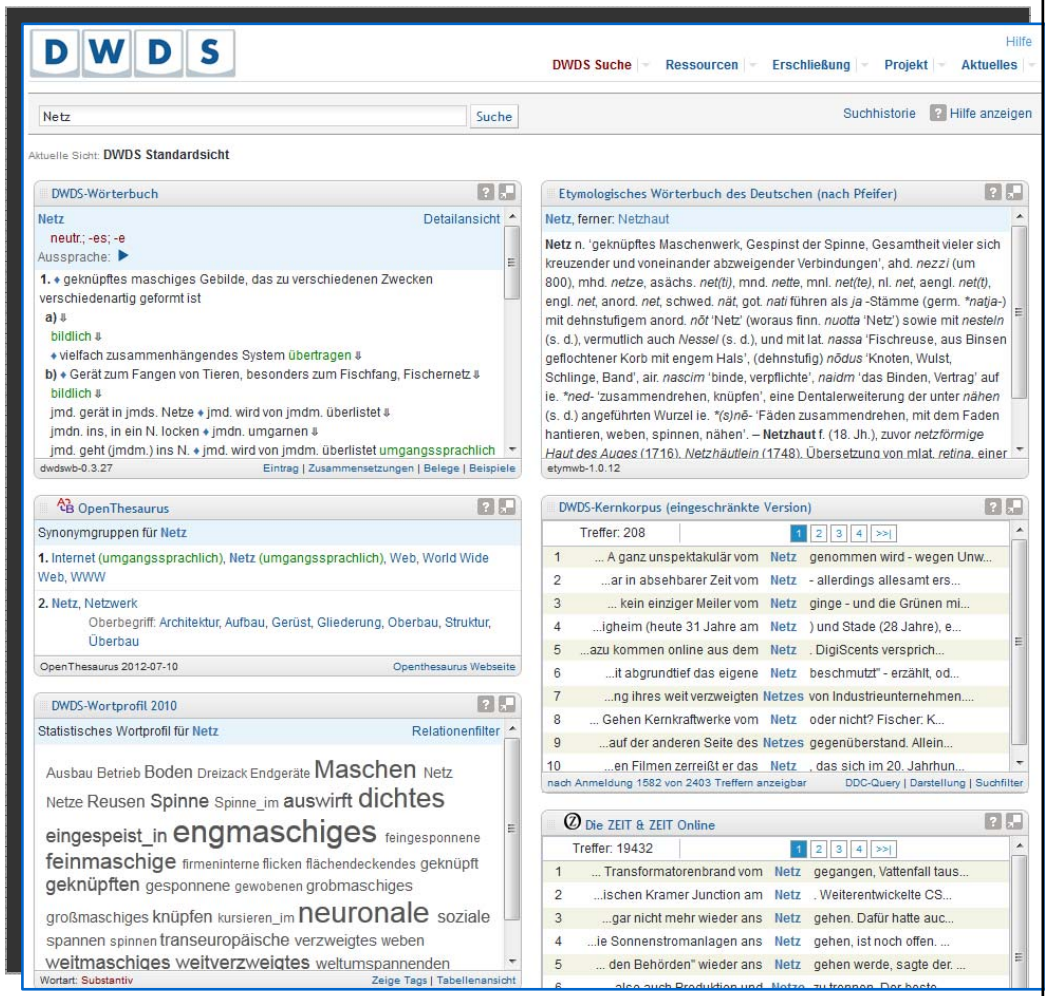
- Berlin-Brandenburgische Akademie der Wissenschaften (BBAW) / DWDS-Projekt (A. Geyken, L. Lemnitzer, M. Ermakova)
- TU Dortmund, Institut für deutsche Sprache und Literatur (M. Beißwenger, A. Storrer)

M. Beißwenger, M. Ermakova, A. Geyken, L. Lemnitzer, A. Storrer (2012): DeRIK – A German Reference Corpus of Computer-Mediated Communication. In: Proceedings of Digital Humanities (DH2012), Hamburg, July 2012.

DeRiK – DWDS

DeRiK ist konzipiert als zusätzliche Komponente des digitalen lexikalischen Informationssystems zum deutschen Wortschatz in Vergangenheit und Gegenwart, das vom Projekt „Digitales Wörterbuch der deutschen Sprache“ (DWDS) unter www.dwds.de online bereitgestellt wird und das die folgenden Ressourcen umfasst:

- Lexikalische Ressourcen (Wörterbücher)
- Korpora (u.a. Referenzkorpus der deutschen Sprache des 20. und frühen 21. Jahrhunderts)
- Statistische Ressourcen (u.a. Wort- und Frequenzprofile)



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

Die Motivation hinter *DeRiK*

Das Kommunikationsaufkommen, das tagtäglich unter Nutzung internetbasierter Kommunikationstechnologien abgewickelt wird, steht derzeit in einem Missverhältnis zur Berücksichtigung des Sprachgebrauchs in internetbasierter Kommunikation (*IBK*) in den Korpora des Gegenwartsdeutschen.

Auch existieren derzeit nur sehr wenige Spezialkorpora zum Sprachgebrauch in der inernetbasierten Kommunikation

– z.B.: **Dortmunder Chat-Korpus**: 1MWord corpus of German chat communication, TU Dortmund, http://www.chatkorpus.tu-dortmund.de

Die Motivation hinter *DeRiK*

	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
14-19 J.	6,3	15,6	30,0	48,5	67,4	76,9	92,1	94,7	95,7	97,3	95,8	97,2	97,5	100,0	100,0	100,0
20-29 J.	13,0	20,7	33,0	54,6	65,5	80,3	81,9	82,8	85,3	87,3	94,3	94,8	95,2	98,4	98,2	98,6
30-39 J.	12,4	18,9	24,5	41,1	50,3	65,6	73,1	75,9	79,9	80,6	81,9	87,9	89,4	89,9	94,4	97,6
40-49 J.	7,7	11,1	19,6	32,2	49,3	47,8	67,4	69,9	71,0	72,0	73,8	77,3	80,2	81,9	90,7	89,4
50-59 J.	3,0	4,4	15,1	22,1	32,2	35,4	48,8	52,7	56,5	60,0	64,2	65,7	67,4	68,9	69,1	76,8
60 +	0,2	0,8	1,9	4,4	8,1	7,8	13,3	14,5	18,4	20,3	25,1	26,4	27,1	28,2	34,5	39,2

Entwicklung der Onlinenutzung in Deutschland 1997 bis 2012: gelegentliche Onlinenutzung
(Quelle: ARD/ZDF-Onlinestudie 2012)

		Gesamt	14-29 J.	30-49 J.	50-69 J.	ab 70 J.
1	Suchmaschinen nutzen	83	96	83	74	58
2	**senden/empfangen von E-Mails**	79	81	83	74	69
3	zielgerichtet bestimmte Angebote/informationen suchen	61	69	64	53	38
4	einfach so im Internet surfen	43	59	40	35	22
5	**Onlinecommunitys nutzen**	36	75	30	12	4
6	Homebanking	32	26	39	28	24
7	**Gesprächsforen/Chatten**	26	56	19	8	4
8	**Instant Messaging**	18	36	15	8	4
9	überspielen/herunterladen von Dateien	17	30	15	10	3
10	Kartenfunktionen nutzen	17	27	15	10	12
11	**Onlinespiele**	16	23	16	7	14
12	"Apps" auf Mobilgeräten nutzen, um ins Internet zu gehen	15	32	11	5	2
13	Musikdateien aus dem Internet	12	29	9	1	
14	Video/TV zeitversetzt	11	20	11	5	2
15	live im Internet Radio hören	10	12	13	7	5
16	RSS-feeds/Newsfeeds	10	15	9	6	1
17	Onlineauktionen/Versteigerungen	7	6	9	5	3
18	andere Waren bestellen/Onlineshopping	7	8	8	5	1
19	live im Internet fernsehen	7	11	7	4	8
20	Buch- und CD-Bestellungen	6	7	6	4	1

Genutzte Onlineanwendungen 2012 (Top 20): mindestens einmal wöchentlich genutzt (Angaben in Prozent) (Quelle: ARD/ZDF-Onlinestudie 2012).

Herausforderungen beim Aufbau eines IBK-Korpus

Gründe für die unbefriedigende Berücksichtigung der internetbasierten Kommunikation in Korpora (u.a.):

- *Gegenstandsbedingte:* Schnelle Veränderung und Weiterentwicklung des Gegenstandes „Internetbasierte Kommunikation“ → Frage der Datenauswahl
- *Juristische:* unklarer rechtlicher Status von IBK-Daten
- *Die Repräsentation/Beschreibung betreffende:*
  Es gibt derzeit keine etablierten Formate (Standards) für die Repräsentation von IBK-Genres und für die Annotation IBK-spezifischer Phänomene
- *Die linguistische Verarbeitung/Aufbereitung betreffende:*
  Gängige NLP-Tools können IBK nicht zufriedenstellend verarbeiten (→ s. Vorträge Beißwenger/Zesch/Evert und Geyken/Jurish/Würzner)

⇨ Kooperation mit empirikom

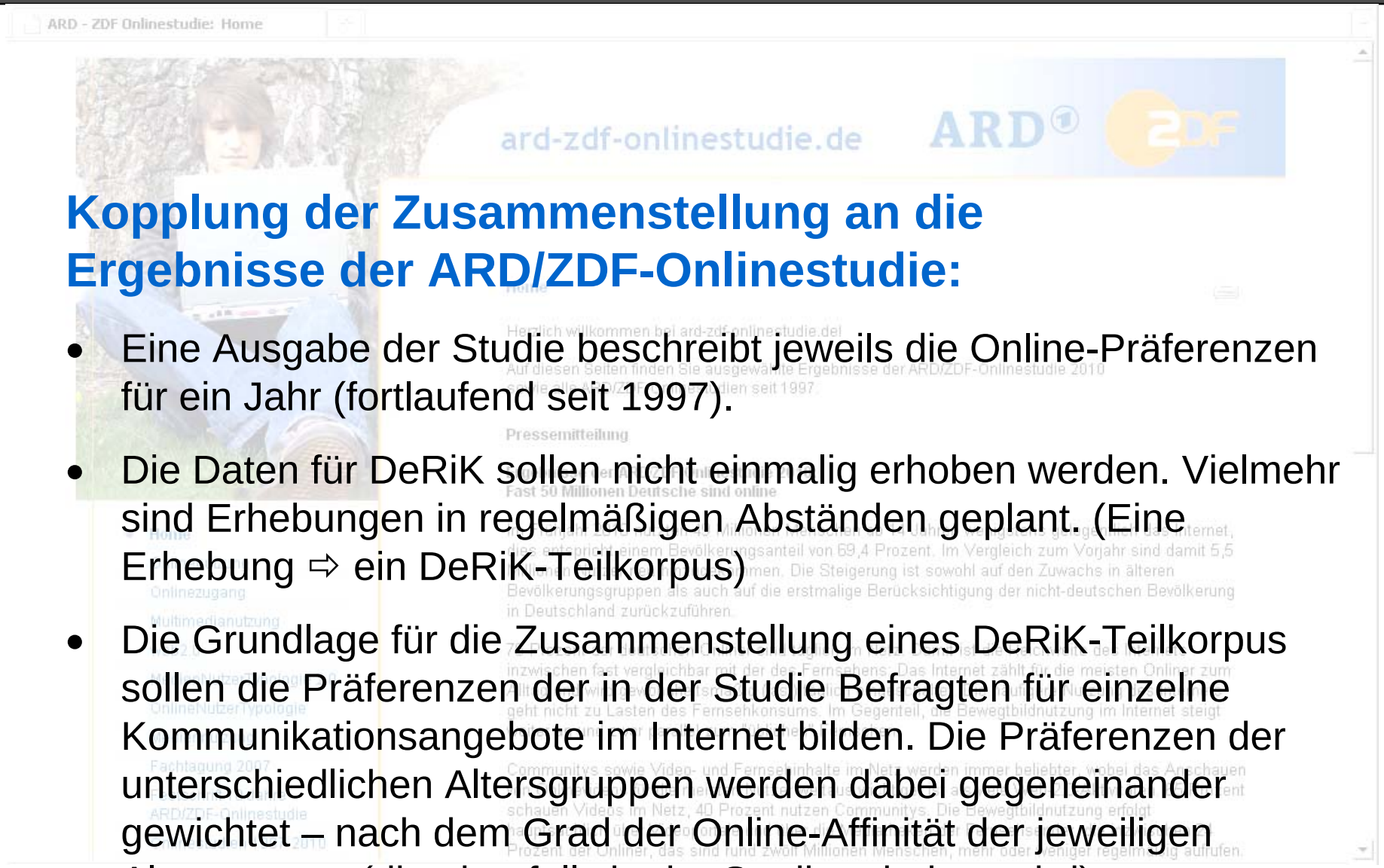
DeRiK: Bisherige Schritte und laufende Arbeiten

- Erabeitung eines *Ideal*konzepts für ein ausgewogenes Korpus zur deutschsprachigen internetbasierten Kommunikation
- Erhebung eines Testkorpus mit Daten aus diversen IBK-Genres (als Testbett für die Arbeiten zur Repräsentation und Verarbeitung) – *112.000 Tokens*
- Manuelle Auswahl erster Datenquellen (die nach derzeitigem Stand juristisch unbedenklich erscheinen)
- Erarbeitung eines Repräsentationsformats für Genres und ausgewählte Phänomene internetbasierter Kommunikation auf Basis des Encoding Frameworks der *Text Encoding Initiative (TEI)*
- Sondierung von Möglichkeiten zur Optimierung der Verarbeitung und linguistischen Annotation von IBK-Daten

DeRiK: Idealkonzept „Ausgewogenes IBK-Korpus“

http://www.ard-zdf-onlinestudie.de/

DeRiK: Idealkonzept „Ausgewogenes IBK-Korpus“

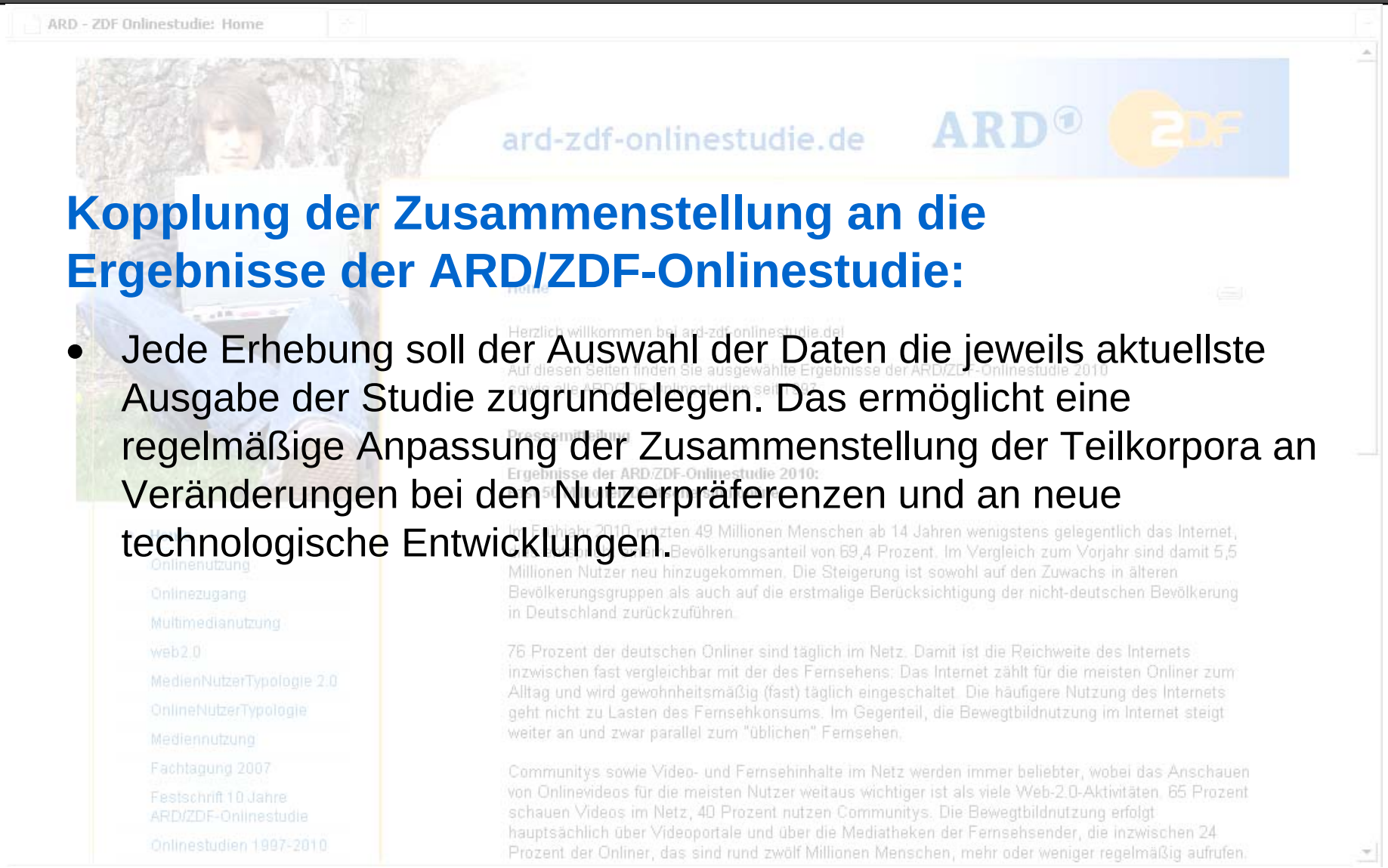


Kopplung der Zusammenstellung an die Ergebnisse der ARD/ZDF-Onlinestudie:

- Eine Ausgabe der Studie beschreibt jeweils die Online-Präferenzen für ein Jahr (fortlaufend seit 1997).
- Die Daten für DeRiK sollen nicht einmalig erhoben werden. Vielmehr sind Erhebungen in regelmäßigen Abständen geplant. (Eine Erhebung ⇨ ein DeRiK-Teilkorpus)
- Die Grundlage für die Zusammenstellung eines DeRiK-Teilkorpus sollen die Präferenzen der in der Studie Befragten für einzelne Kommunikationsangebote im Internet bilden. Die Präferenzen der unterschiedlichen Altersgruppen werden dabei gegeneinander gewichtet – nach dem Grad der Online-Affinität der jeweiligen Altersgruppe (die ebenfalls in der Studie erhoben wird).

http://www.ard-zdf-onlinestudie.de/

DeRiK: Idealkonzept „Ausgewogenes IBK-Korpus“



Kopplung der Zusammenstellung an die Ergebnisse der ARD/ZDF-Onlinestudie:

- Jede Erhebung soll der Auswahl der Daten die jeweils aktuellste Ausgabe der Studie zugrundelegen. Das ermöglicht eine regelmäßige Anpassung der Zusammenstellung der Teilkorpora an Veränderungen bei den Nutzerpräferenzen und an neue technologische Entwicklungen.

http://www.ard-zdf-onlinestudie.de/

Frage(n) der Datenauswahl

Gründe für die unbefriedigende Berücksichtigung der internetbasierten Kommunikation in Korpora (u.a.):

- *Gegenstandsbedingte:* Schnelle Veränderung und Weiterentwicklung des Gegenstandes „Internetbasierte Kommunikation“ → Frage der Datenauswahl
- *Juristische:* unklarer rechtlicher Status von IBK-Daten

⇨ Die Kopplung der Datenerhebung an die ARD/ZDF-Studie ist ein Idealkonzept. So lange die Rechtslage zur Verwendung von IBK-Daten für Forschungszwecke und für die Nutzung in Korpora unklar ist, kann dieses Konzept nur so weit umgesetzt werden, als sich Daten aus den relevanten Genres finden lassen, die explizit unter einer CC-Lizenz stehen. Ändert sich die Rechtslage, können im Zuge weiterer Erhebungen Anpassungen vorgenommen werden.

Herausforderungen beim Aufbau eines IBK-Korpus

Gründe für die unbefriedigende Berücksichtigung der internetbasierten Kommunikation n Korpora (u.a.):

- *Gegenstandsbedingte:* Schnelle Veränderung und Weiterentwicklung des Gegenstandes „Internetbasierte Kommunikation“ → Frage der Datenauswahl
- *Juristische:* unklarer rechtlicher Status von IBK-Daten
- *Die Repräsentation/Beschreibung betreffende:*
  Es gibt derzeit keine etablierten Formate (Standards) für die Repräsentation von IBK-Genres und für die Annotation IBK-spezifischer Phänomene
- *Die linguistische Verarbeitung/Aufbereitung betreffende:*
  Gängige NLP-Tools können IBK nicht zufriedenstellend verarbeiten (→ s. Vorträge Beißwenger/Zesch/Evert und Geyken/Jurish/Würzner)

Towards a (TEI) target format for encoding CMC

- As the DWDS corpus is annotated by the means of TEI (Text Encoding Initiative), we are going to use TEI for encoding DeRiK as well.
- It is necessary to customize the TEI guidelines for the needs of CMC annotation.

  (CMC data in its breadth does not fit either the „written text“ <p> nor the „spoken language“ <u> model of the TEI guidelines, it is dialogic but in a written medium)
- The customized, TEI-compliant annotation schema which we have developed so far provides:
  - ➢ a model for the macrostructure for CMC discourse
  - ➢ suggestion for the description of CMC-specific phenomena

M. Beißwenger, M. Ermakova, A. Geyken, L. Lemnitzer, A. Storrer (2012, im Druck):
A TEI Schema for the Representation of Computer-Mediated Communication.
In: *Journal of the Text Encoding Initiative (TEI)*.

< Text Encoding Initiative >

Macrostructure vs. microstructure

Macrostructure

describes how the postings are sequenced.

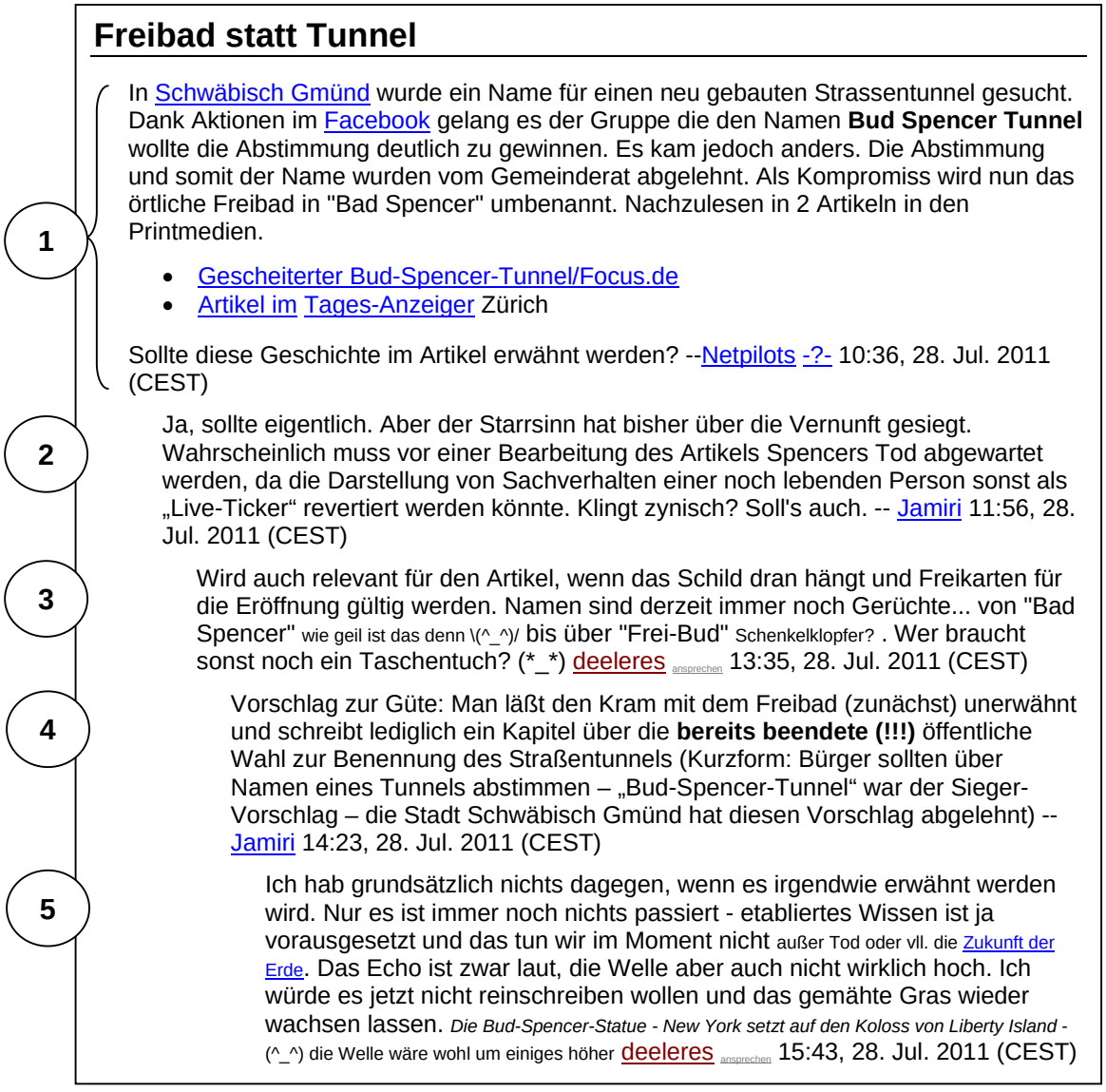
Microstructure

refers to the internal structure of posting.

The element „posting“

Posting

New category/element ***<posting>*** as a basic element and pivot between macrostructure and microstructure.



Example (taken from ex. 5 in the handout)

```
<div type=„thread“>
    <posting indentLevel="1" synch="#t02"
      who="#A02">
      <p>Weil man dass dann auch so darstellen
      sollte mit diesem Link und nicht mit dem
      hinweis auf die Comics... --
      <autoSignature/></p>
    </posting>
    <posting indentLevel="3" synch="#t04"
      who="#A04">
      <p>Per Umweltschützen. Das hat in dem
      Artikwel aber auch mal gar nichts zu suchen!
      --<autoSignature/></p>
    </posting>
</div>
```

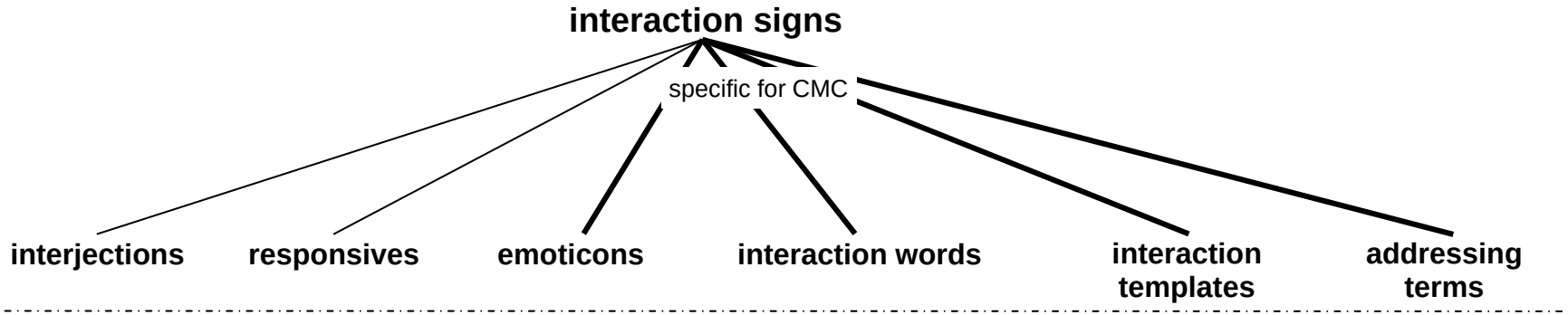
Microstructure – interaction signs

The concept „**interaction sign**“ adopts the concept „**Interaktive Einheiten**“ from the GDS (Zifonun/Hoffmann/Strecker 1997) and encompasses

- *interjections* („hm“, „oh my god“)
- *responsives* („yes“, „no“) as well as
- some new classes which are typical for CMC:
  - *Emoticons*
  - *interaction words*
  - *interaction templates*
  - *addressing terms*

See also examples 3 to 5 in the handout

A taxonomy of interaction signs



EXAMPLES

interjections	responsives	emoticons	interaction words	interaction templates	addressing terms
ach äh eiei mhm ne oho tja äh au gell na oh oi (etc.) ah aua hm naja oh pst	ja okay nein	*Western style:* :-) :) ;-) ;) :-(:(:-D :D *Japanese style:* o.O O.O \(^_^)/ (*_*)	grins freu ganzdolleknuddel grübel schluck stirnrunzel lach schluchz lol rofl	Ok Exzellent	@zora: … @tinchen: … an bilbo21: …

Where are we now?

- We have released a stable, TEI-conformant schema and documentation, the schema is open to further modifications: http://www.empirikom.net/bin/view/Themen/CmcTEI
- Publication: 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 TEI* (in press).
- We have acquired samples of some CMC genres: Wikipedia discussions, blogs and forums, Twitter data, chat data via ‚Dortmunder Chatkorpus'.
- We have provided TEI metadata for these documents.

DeRiK: Nächste Schritte

- Test des TEI-Repräsentationsformats mit weiteren Daten und Genres (bisher im Fokus. v.a. Wikipedia-Diskussionen, Chat/Instant Messaging-Logfiles, Foren-Threads) und ggf. Verfeinerung
- Konvertierung der Daten in TEI
- Experimente zur Verarbeitung und linguistischen Annotation von IBK-Daten (in Kooperation mit anderen Akteuren im Empirikom-Netzwerk)
- Festlegung einer Zielgröße und von Erhebungszeitpunkten für das DeRiK-Korpus
- Festlegung der Zusammenstellung für das erste Teilkorpus
  ⇨ Erhebung der Daten
  ⇨ TEI-Annotation
  (⇨ Linguistische Annotation)

Aufbau und Annotation eines Referenzkorpus zur deutsch-sprachigen internetbasierten Kommunikation (*DeRiK*-Projekt)

technische universität dortmund

Michael Beißwenger

berlin-brandenburgische AKADEMIE DER WISSENSCHAFTEN

Lothar Lemnitzer

Workshop: Webkorpora in Computerlinguistik und Sprachforschung // Mannheim, 27./28. Sept. 2012

GSCL

INSTITUT FÜR DEUTSCHE SPRACHE