

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

tu 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)



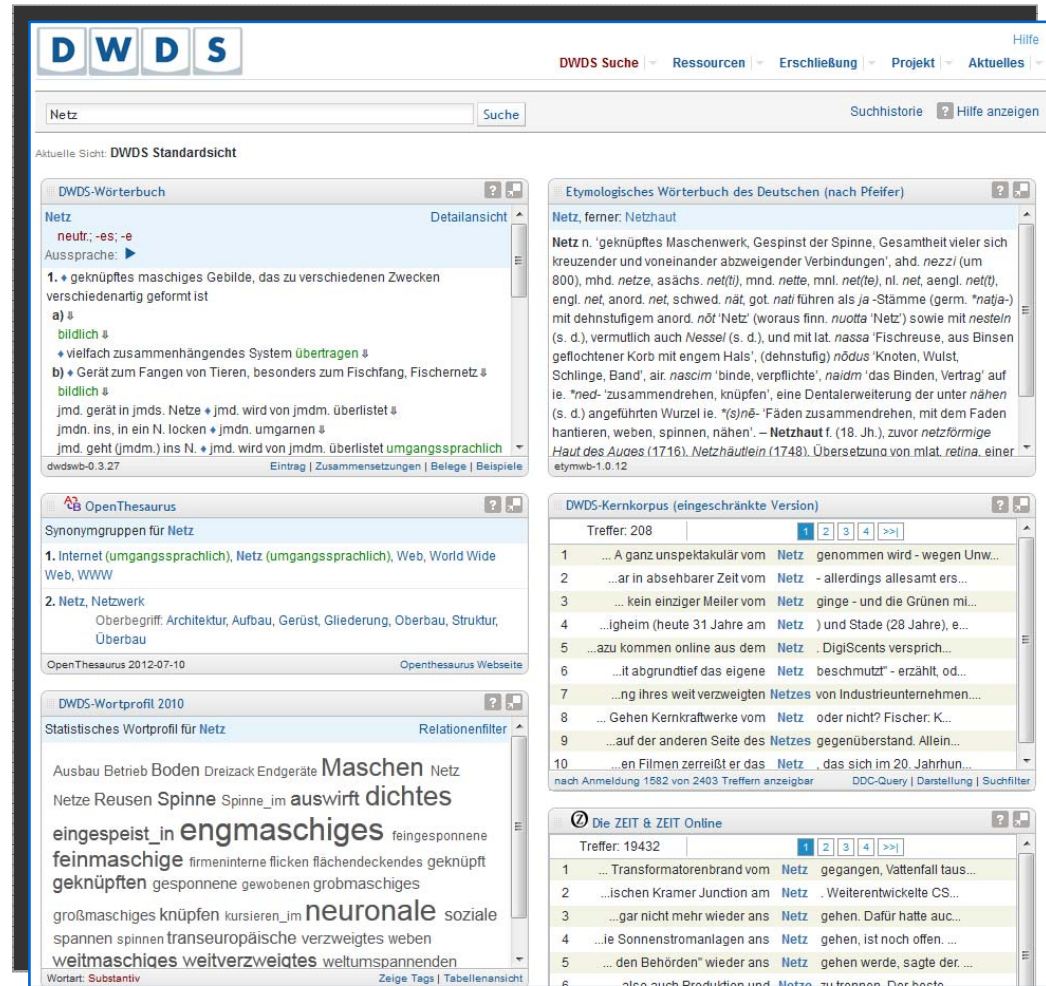
- TU Dortmund, Institut für deutsche Sprache und Literatur (M. Beißwenger, A. Storrer)



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 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)



The screenshot shows the DWDS website interface. At the top, there are navigation tabs for 'DWDS Suche', 'Ressourcen', 'Erschließung', 'Projekt', and 'Aktuelles'. A search bar contains the word 'Netz'. Below the search bar, the 'Aktuelle Sicht: DWDS Standardsicht' is displayed. The main content area is divided into several panels:

- DWDS-Wörterbuch:** Shows the entry for 'Netz' with its grammatical information (neutr.; -es; -e), pronunciation, and a definition: 'geknüpftes maschiges Gebilde, das zu verschiedenen Zwecken verschiedenartig geformt ist'. It also lists related terms like 'übertragen' and 'umgangssprachlich'.
- Etymologisches Wörterbuch des Deutschen (nach Pfeifer):** Provides a detailed etymology of 'Netz', tracing it back to Old High German and Old Norse roots, and mentioning related terms like 'Netzhaut' and 'retina'.
- OpenThesaurus:** Shows synonym groups for 'Netz', including 'Internet', 'Netzwerk', and 'Web'.
- DWDS-Wortprofil 2010:** Displays a statistical word profile for 'Netz' with various related terms and their frequencies.
- DWDS-Kernkorpus (eingeschränkte Version):** Shows a list of 208 hits for the word 'Netz' in the corpus, with the first few entries visible.
- Die ZEIT & ZEIT Online:** Shows a list of 19432 hits for 'Netz' in the newspaper corpus, with the first few entries visible.

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.

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 internetbasierten Kommunikation

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

	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).

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*

- Erarbeitung eines *Idealkonzepts* 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

ARD - ZDF Onlinestudie: Home



ard-zdf-onlinestudie.de **ARD** **ZDF**

Home 

Herzlich willkommen bei ard-zdf-onlinestudie.de!

Auf diesen Seiten finden Sie ausgewählte Ergebnisse der ARD/ZDF-Onlinestudie 2010 sowie alle ARD/ZDF-Onlinestudien seit 1997.

Pressemitteilung

**Ergebnisse der ARD/ZDF-Onlinestudie 2010:
Fast 50 Millionen Deutsche sind online**

Im Frühjahr 2010 nutzten 49 Millionen Menschen ab 14 Jahren wenigstens gelegentlich das Internet, dies entspricht einem Bevölkerungsanteil von 69,4 Prozent. Im Vergleich zum Vorjahr sind damit 5,5 Millionen Nutzer neu hinzugekommen. Die Steigerung ist sowohl auf den Zuwachs in älteren Bevölkerungsgruppen als auch auf die erstmalige Berücksichtigung der nicht-deutschen Bevölkerung in Deutschland zurückzuführen.

76 Prozent der deutschen Onliner sind täglich im Netz. Damit ist die Reichweite des Internets inzwischen fast vergleichbar mit der des Fernsehens: Das Internet zählt für die meisten Onliner zum Alltag und wird gewohnheitsmäßig (fast) täglich eingeschaltet. Die häufigere Nutzung des Internets geht nicht zu Lasten des Fernsehkonsums. Im Gegenteil, die Bewegtbildnutzung im Internet steigt weiter an und zwar parallel zum "üblichen" Fernsehen.

Communitys sowie Video- und Fernsehinhalte im Netz werden immer beliebter, wobei das Anschauen von Onlinevideos für die meisten Nutzer weitaus wichtiger ist als viele Web-2.0-Aktivitäten. 65 Prozent schauen Videos im Netz, 40 Prozent nutzen Communitys. Die Bewegtbildnutzung erfolgt hauptsächlich über Videoportale und über die Mediatheken der Fernsehsender, die inzwischen 24 Prozent der Onliner, das sind rund zwölf Millionen Menschen, mehr oder weniger regelmäßig aufrufen.

- **Home**
- Onlinenutzung
- Onlinezugang
- Multimedienutzung
- web2.0
- MedienNutzerTypologie 2.0
- OnlineNutzerTypologie
- Mediennutzung
- Fachtagung 2007
- Festschrift 10 Jahre ARD/ZDF-Onlinestudie
- Onlinestudien 1997-2010

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).

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.

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.

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)*.

Macrostructure

describes how the postings are sequenced.

`<div type=„logfile“>` `<div type=„thread“>`

Microstructure

refers to the internal structure of posting.

Posting

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

Freibad statt Tunnel

1 In [Schwäbisch Gmünd](#) wurde ein Name für einen neu gebauten Strassentunnel gesucht. Dank Aktionen im [Facebook](#) gelang es der Gruppe die den Namen **Bud Spencer Tunnel** wollte die Abstimmung deutlich zu gewinnen. Es kam jedoch anders. Die Abstimmung und somit der Name wurden vom Gemeinderat abgelehnt. Als Kompromiss wird nun das örtliche Freibad in "Bad Spencer" umbenannt. Nachzulesen in 2 Artikeln in den Printmedien.

- [Gescheiterter Bud-Spencer-Tunnel/Focus.de](#)
- [Artikel im Tages-Anzeiger](#) Zürich

Sollte diese Geschichte im Artikel erwähnt werden? --[Netpilots](#) -?- 10:36, 28. Jul. 2011 (CEST)

2 Ja, sollte eigentlich. Aber der Starrsinn hat bisher über die Vernunft gesiegt. Wahrscheinlich muss vor einer Bearbeitung des Artikels Spencers Tod abgewartet werden, da die Darstellung von Sachverhalten einer noch lebenden Person sonst als „Live-Ticker“ revertiert werden könnte. Klingt zynisch? Soll's auch. -- [Jamiri](#) 11:56, 28. Jul. 2011 (CEST)

3 Wird auch relevant für den Artikel, wenn das Schild dran hängt und Freikarten für die Eröffnung gültig werden. Namen sind derzeit immer noch Gerüchte... von "Bad Spencer" wie geil ist das denn \(^_\^)/ bis über "Frei-Bud" Schenkelklopfer? . Wer braucht sonst noch ein Taschentuch? (*_*) [deeleres](#) ansprechen 13:35, 28. Jul. 2011 (CEST)

4 Vorschlag zur Güte: Man läßt den Kram mit dem Freibad (zunächst) unerwähnt und schreibt lediglich ein Kapitel über die **bereits beendete (!!!)** öffentliche Wahl zur Benennung des Straßentunnels (Kurzform: Bürger sollten über Namen eines Tunnels abstimmen – „Bud-Spencer-Tunnel“ war der Sieger-Vorschlag – die Stadt Schwäbisch Gmünd hat diesen Vorschlag abgelehnt) -- [Jamiri](#) 14:23, 28. Jul. 2011 (CEST)

5 Ich hab grundsätzlich nichts dagegen, wenn es irgendwie erwähnt werden wird. Nur es ist immer noch nichts passiert - etabliertes Wissen ist ja vorausgesetzt und das tun wir im Moment nicht außer Tod oder vll. die [Zukunft der Erde](#). Das Echo ist zwar laut, die Welle aber auch nicht wirklich hoch. Ich würde es jetzt nicht reinschreiben wollen und das gemähte Gras wieder wachsen lassen. *Die Bud-Spencer-Statue - New York setzt auf den Koloss von Liberty Island -* (^_\^) die Welle wäre wohl um einiges höher [deeleres](#) ansprechen 15:43, 28. Jul. 2011 (CEST)

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>
```

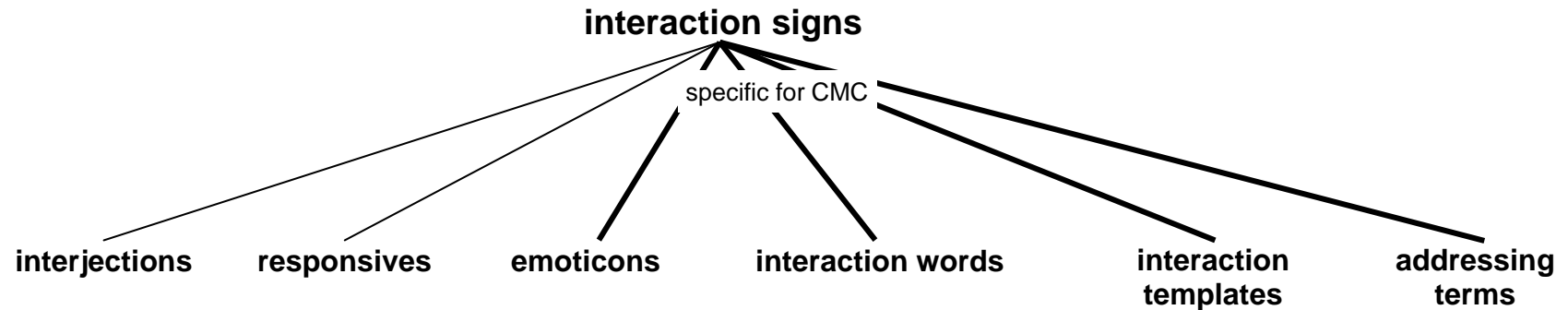
```
</div>
```

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

ach äh ah
 äh au aua
 eiei gell hm
 mhm na naja
 ne oh oh
 oho oi pst
 tja (etc.)

ja
 okay
 nein

Western style:

:-)
 ;-)
 :-(
 :-D

Japanese style:

o.O O.O
 \(^_^)/ (*_*)

grins freu
 ganzdolleknuddel
 grübel schluck
 stirnrunzel
 lach schluchz

lol rofl



@zora: ...
 @tinchen: ...
 an bilbo21: ...

- 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.

- 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)

tu 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