Building a Reference Corpus of German Computer-Mediated Communication (DeRiK-Project)

Michael Beißwenger
Angelika Storrer

Alexander Geyken
Maria Ermakova
Lothar Lemnitzer

Workshop: Building Corpora of Computer-Mediated Communication: Issues, Challenges, Perspectives
Dortmund, 14. Feb. 2013
DeRiK

„Deutsches Referenzkorpus zur internetbasierten Kommunikation“ (DeRiK)

**Aims:** Building, Maintaining and Annotating a Reference Corpus of German CMC

**Cooperation Partners:**

- Berlin-Brandenburgische Akademie der Wissenschaften (BBAW) / DWDS-Projekt
- TU Dortmund, Institut für deutsche Sprache und Literatur

DeRiK should serve as a component of a digital lexical system that documents the use of the German language past and present. The DWDS digital lexical system (www.dwds.de) offers one-click-access to three types of resources:

- Lexical (dictionaries)
- Various Corpora including a Core Corpus (>1900)
- Lexical Statistics

The CMC corpus will be an additional resource.

Up to now, there are only very few specialized corpora, e.g.

- the **Dortmund Chat Corpus** (1MWord corpus of German chat communication, TU Dortmund, http://www.chatkorpus.tu-dortmund.de)

- a corpus of German tweets

Also the large balanced corpora of contemporary language do not include CMC.
Up to now, there are only very few specialized corpora, e.g.

• the **Dortmund Chat Corpus** (1MWord corpus of German chat communication, TU Dortmund, http://www.chatkorpus.tu-dortmund.de)

• a corpus of German tweets

Also the large balanced corpora of contemporary language do not include CMC.
A large and balanced corpus of CMC is a desideratum for several fields of research, e.g.:

- empirical research on the peculiarities of language use on the internet (= linguistic cmc research in the narrower sense);
- empirical linguistic analysis, documentation and lexicographic description of current trends in contemporary language;
- for corpus-based research on the impact of cmc on language change.
## Entwicklung der Onlinenutzung in Deutschland 1997 bis 2012: gelegentliche Onlinenutzung

(Quelle: ARD/ZDF-Onlinestudie 2012).

### Motivation

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>14-19 J.</td>
<td>6,3</td>
<td>15,6</td>
<td>30,0</td>
<td>48,5</td>
<td>67,4</td>
<td>76,9</td>
<td>92,1</td>
<td>94,7</td>
<td>95,7</td>
<td>97,3</td>
<td>95,8</td>
<td>97,2</td>
<td>97,5</td>
<td>100,0</td>
<td>100,0</td>
<td>100,0</td>
</tr>
<tr>
<td>20-29 J.</td>
<td>13,0</td>
<td>20,7</td>
<td>33,0</td>
<td>54,6</td>
<td>65,5</td>
<td>80,3</td>
<td>81,9</td>
<td>82,8</td>
<td>85,3</td>
<td>87,3</td>
<td>94,3</td>
<td>94,8</td>
<td>95,2</td>
<td>98,4</td>
<td>98,2</td>
<td>98,6</td>
</tr>
<tr>
<td>30-39 J.</td>
<td>12,4</td>
<td>18,9</td>
<td>24,5</td>
<td>41,1</td>
<td>50,3</td>
<td>65,6</td>
<td>73,1</td>
<td>75,9</td>
<td>79,9</td>
<td>80,6</td>
<td>81,9</td>
<td>87,9</td>
<td>89,4</td>
<td>89,9</td>
<td>94,4</td>
<td>97,6</td>
</tr>
<tr>
<td>40-49 J.</td>
<td>7,7</td>
<td>11,1</td>
<td>19,6</td>
<td>32,2</td>
<td>49,3</td>
<td>47,8</td>
<td>67,4</td>
<td>69,9</td>
<td>71,0</td>
<td>72,0</td>
<td>73,8</td>
<td>77,3</td>
<td>80,2</td>
<td>81,9</td>
<td>90,7</td>
<td>89,4</td>
</tr>
<tr>
<td>50-59 J.</td>
<td>3,0</td>
<td>4,4</td>
<td>15,1</td>
<td>22,1</td>
<td>32,2</td>
<td>35,4</td>
<td>48,8</td>
<td>52,7</td>
<td>56,5</td>
<td>60,0</td>
<td>64,2</td>
<td>65,7</td>
<td>67,4</td>
<td>68,9</td>
<td>69,1</td>
<td>76,8</td>
</tr>
<tr>
<td>60 +</td>
<td>0,2</td>
<td>0,8</td>
<td>1,9</td>
<td>4,4</td>
<td>8,1</td>
<td>7,8</td>
<td>13,3</td>
<td>14,5</td>
<td>18,4</td>
<td>20,3</td>
<td>25,1</td>
<td>26,4</td>
<td>27,1</td>
<td>28,2</td>
<td>34,5</td>
<td>39,2</td>
</tr>
</tbody>
</table>

### Genutzte Onlineanwendungen 2012 (Top 20): mindestens einmal wöchentlich genutzt


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

Reasons why CMC genres are still not well represented in contemporary language corpora

• **Relating to the subject:** The fast and constant change of channels, media, modes of CMC makes ‘CMC’ a moving target → influence on data collection

• **Relating to legal and ethical aspects:** ownership, persistence of data; issues of anonymization

• **Wrt the representation of the primary data:** there are (not yet) established standards for the representation and annotation of such data

• **Relating to linguistic processing and annotation:** NLP tools as well as annotation tools have to be adapted to “non-standard” ways of writing.
DeRiK: what has been done so far

- Developing a concept of the ‘ideal’ balanced CMC corpus
- Acquisition of a sample corpus of various CMC genres (wikipedia talk pages, forums, tweets, chats) ~500,000 tokens
- Selection of data sources according to the legal status of the data (explicit use of distribution licences)
- Shaping of a representation format for the data as a customization of the Text Encoding Initiative (TEI)
- Adaptation of tools and manual annotation of data (training corpus)
The data sampling process and the “ARD/ZDF-Onlinestudie”

- The annual “online study” is a survey of the use of online media across different age groups (data available since 1997).
- Ideally, the survey should inform the selection and sampling of CMC genres in such a reference corpus.
- Controlling variables like the age of the users depends on the availability of metadata.

http://www.ard-zdf-onlinestudie.de/
The data sampling process and the “ARD/ZDF-Onlinestudie”

- Each sampling period should take into account the latest survey, in order not to miss emerging new media and tools of communication (twitter, facebook ...)

⇒ The data sampling process must take into account legal and ethical issues of the use and distribution of data. Currently, we therefore have to follow a more opportunistic approach to data sampling (data which explicitly use open access and open content licences). Once there is a stable workflow, additional data can be acquired and added in order to arrive at a balanced data sample
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 details will be layed out in the talk of Beißwenger and Ermakova

DeRiK: Current work

- A TEI-style customized schema has been developed and is available for the community
- We are currently developing converters from various source formats (e.g. WP talk page) to the target format
- This includes the acquisition of metadata in a metadata header and the anonymization of the contributors
- Major challenges are the “up”-conversion from an informationally poor, layout-oriented markup to an informationally richer markup and a reasonable solution of ambiguities of the original format

(Work by: Maria Ermakova and Frank Wiegand)
Linguistic analysis and annotation

- Use of statistical segmenter/tokeniser and part-of-speech tagger (tools which have been developed at the BBAW)
- The tools have already been used for other non-standard variants of German (diachronic corpora)
- Manual segmentation and POS-annotation of a training corpus (WP talk pages, chats, forums, tweets, at least 50,000 tokens)

Training of the POS-tagger with a slightly modified version of the Stuttgart-Tübingen tag set

(Work by: Kay-Michael Würzner, Gabriella Pein and Bryan Jurish)
Example: segmentation

Tasks: Tokenization and Sentence Boundary Detection

Motivation:
Our tools must handle German „standard“ as well as different kinds of „non-standard“ formats, e.g.
- Computer-mediated communication
- Diachronic corpora of various language stages

Approach: A „one does it all“ segmenter with
- Statistical approach to the task
- Different kinds of training and gold-standard corpora for the different types of data, i.e. segmentation tasks
Example: segmentation

Current state with „standard“ data

• Reference „gold-standard“: TIGER corpus of written contemporary German (newspaper texts, ~ 1 mill tokens)
• Accuracy achieved for the combined tasks of SBD and tokenisation: 99.8 %, where tokenization is more accurate than SBD
• Approach: SBD as detection of probable sentence starts (in contrast to e.g. Kiss/Strunk 2006)

Dealing with non standard-data

• Building a segmented/annotated corpus as a proof of concept of our approach
• Developing guidelines for the annotation of sentence boundaries for building a gold-standard corpus
Example: segmentation

Challenges

• Sentence boundaries are frequently not marked by full-stop
• Sentences cross the border of turns (in chat data)
• Approach: each token boundary is ambiguous (first word of a sentence – yes or no?)
• Look for a local maximum of the probability that a given token boundary is also a sentence-boundary (using a 2nd-order HMM, i.e. trigrams)
Thank you…

… for your attention.

Special thanks to

Gabriella Pein for annotating the reference corpus for the tokenisation and POS-tagging task

Kay-Michael Würzner for the adaptation of the annotation tools

For questions, please contact:

Maria Ermakova: ermakovamd@googlemail.com for details of the TEI Schema

Michael Beißwenger and Angelika Storrer
{beisswenger, storrer}@uni-dortmund.de for linguistic aspects

Kay-Michael Würzner and Lothar Lemnitzer
{wuerzner, lemnitzer}@bbaw.de for processing aspects