DeRiK: A German reference corpus of computer-mediated communication

Michael Beißwenger
TU Dortmund University, Institut für deutsche Sprache und Literatur, Germany

Maria Ermakova, Alexander Geyken and Lothar Lemnitzer
Berlin-Brandenburg Academy of Sciences and Humanities, Digitales Wörterbuch der deutschen Sprache, Germany

Angelika Storrer
TU Dortmund University, Institut für deutsche Sprache und Literatur, Germany

Abstract

The article describes an ongoing project that aims at building a reference corpus of German computer-mediated communication (CMC) as a new component of an already existing reference corpus of written contemporary German. The 'Deutsches Referenzkorpus zur internetbasierten Kommunikation' (*DeRiK*) shall include data from the most prominent CMC genres amongst German Internet users and, thus, close a gap in the coverage of the corpus resources in the project 'Digitales Wörterbuch der deutschen Sprache' (DWDS), which are maintained and provided by the Berlin-Brandenburg Academy of Sciences and Humanities. The focus of the article is on the role of the DeRiK component within the DWDS framework, on sampling issues, and on CMC-specific issues of corpus annotation.

Correspondence:
Michael Beißwenger, TU Dortmund University, Institut für deutsche Sprache und Literatur, D-44221 Dortmund, Germany
E-mail:
michael.beisswenger@tu-dortmund.de

1 Project Background and Focus of the Article

In view of the increasing amount of reading and writing that people do on the Internet, up-to-date corpora of written contemporary language must take into consideration the impact of computer-mediated communication (CMC) on contemporary language and, thus, include samples of emerging written genres such as e-mail, weblogs, microblogging on Twitter, discussion boards and wiki discussions, chats and instant messaging conversations, and communication in social network sites. In this article, we present selected aspects of an ongoing project that aims at building a reference corpus of German CMC, called 'DeRiK' ('Deutsches Referenzkorpus zur internetbasierten Kommunikation').[1] DeRiK is a joint initiative of TU Dortmund University and the Berlin-Brandenburg Academy of Sciences and Humanities (BBAW) and is embedded in the scientific network 'Empirical Research on Internet-based

Literary and Linguistic Computing, Vol. 28, No. 4, 2013.

doi:10.1093/llc/fqt038 Advance Access published on 3 July 2013

Communication'[2] funded by the Deutsche Forschungsgemeinschaft (DFG). The corpus will be integrated into the lexical information system provided by the BBAW project 'Digitales Wörterbuch der deutschen Sprache' (DWDS, www.dwds.de).

The focus of this article is on the role of the DeRiK component within the DWDS framework and on sampling issues (Section 3) as well as on CMC-specific issues of corpus annotation (Section 4).

2 Related Work

Up to now, there are few corpora of CMC. An overview of CMC corpora as of 2008 is given in Beißwenger and Storrer (2008). Examples for existing corpora are the 'Dortmund Chat Corpus'[3] for language use and linguistic variation in German chats, which was collected and annotated in 2002–8 (Beißwenger, 2013), the 'Queer Chat-Room Corpus' recorded in 2005 (King, 2009), the English 'NPS Chat Corpus'[4] collected and annotated in 2006–8 (Forsyth and Martell, 2007), or the 'Netlog Corpus' for Flemish Dutch Internet Language (Kestemont *et al.*, 2012).

An example for a reference corpus of written language, which—similar to the DWDS/DeRiK project—aims to include a CMC subcorpus is the 'SoNaR' project for contemporary Dutch (Reynaert *et al.*, 2010). Other initiatives and projects, which are currently building reference or specialized corpora on CMC, are, for example, the projects gathered in the scientific network 'Building and Annotating Corpora of Computer-Mediated Communication'[5] and in the French special interest group 'Nouvelles formes des communication (Nouv-com)'.[6]

3 Integrating CMC Discourse into a Corpus of Contemporary German: Motivation, Sampling and Application Fields

DWDS (www.dwds.de) is a lexical information system developed by and hosted at the BBAW. The system offers one-click-access to three different types of resources (Geyken, 2007; Klein and Geyken, 2010):

(a) lexical resources: a common language dictionary,[7] an etymological dictionary, and a thesaurus;
(b) corpus resources: a balanced reference corpus (called 'DWDS core corpus') of German ranging from 1900 up to now, a set of additional newspaper corpora, and specialized corpora;
(c) statistical resources for words and word combinations.

These resources are displayed alongside one another in separate panels (see Fig. 1). The system offers the choice among several views, i.e. between several profiles with predefined panel combinations.

The CMC component *DeRiK* ('Deutsches Referenzkorpus zur internetbasierten Kommunikation') will be integrated into this framework both as an independent panel and as a subcorpus of the DWDS core corpus. The data for DeRiK shall be collected not only once but on a regular basis; DeRiK, thus, will consist of several partial corpora, each of them representing data that have been collected at a certain point of time (e.g. within 1 year).

The focus of the corpus is on 'internet-based' genres of CMC, which are based on the infrastructure of the internet. This makes it possible to link up the selection of relevant genres and the sampling of the data to the results of the annual issues of the 'ARD/ZDF-Onlinestudie', a German online usage survey (www.ard-zdf-onlinestudie.de), which monitors the usage preferences of German Internet users on an annual basis and according to online applications and age groups. The findings of this survey allow us to derive an ideal key for the composition of the DeRiK partial corpora, i.e. for deciding which CMC technologies have to be regarded as most prominent amongst German Internet users in any year and in which proportion discourse conducted on the basis of those technologies should be represented in the corpus. However, for practical reasons, the project will set out to collect data of only those instances of CMC technologies indicated by the online survey for which the users have explicitly granted permission for (re-)distributing and (re-)using their written utterances for non-commercial purposes/academic research (e.g. by assigning the respective subtypes of the 'Creative Commons' License to CMC documents or to

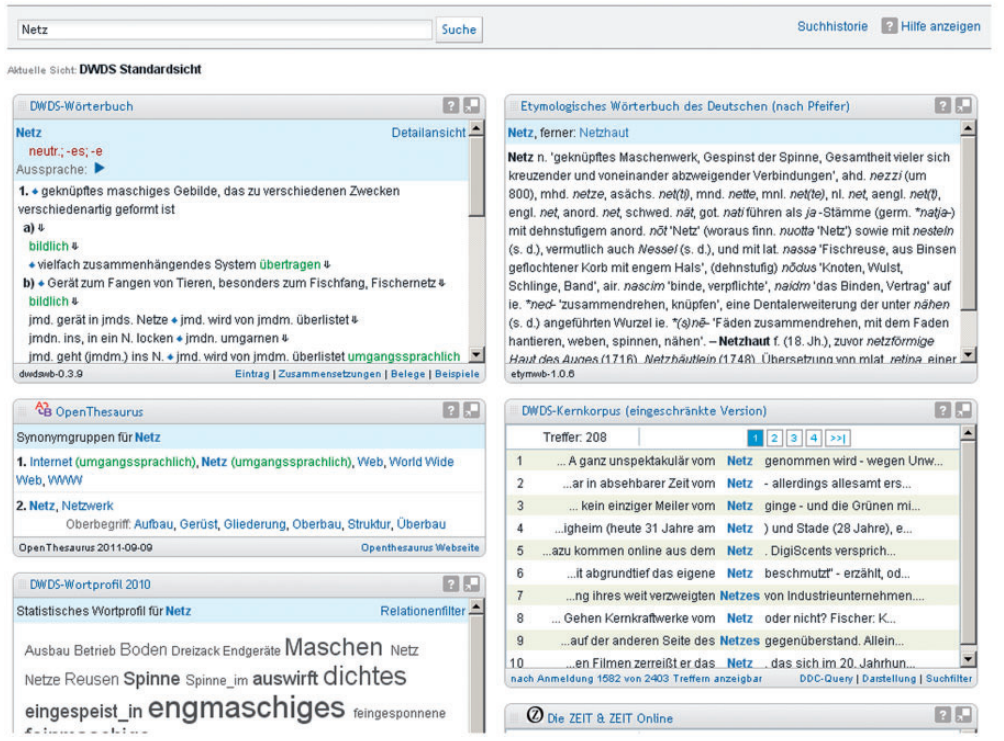

Fig. 1 Web frontend of the DWDS system (http://www.dwds.de)

CMC applications on the web). Thus, the key derived from the findings of the annual online survey will describe an 'ideal' compilation (with ideal proportions of the CMC genres) while the legal constraints will compel us to implement this ideal key only in modified form. As the data will be collected during several years, we will have the possibility to adapt our key for each phase of data collection—to changing usage preferences according to the most recent version of the online survey as well as to changes in restrictions concerning Intellectual Property Rights on the use of CMC data retrieved from the web for scientific purposes.

The first partial corpus of DeRiK will include mostly discourse from Wikipedia talk pages, a selection of forum and weblog discussions, chat conversations, and postings of selected Twitter users who have published their tweets under a Creative Commons license.

It is planned to collect up to 10 million tokens per decade and to automatize as much of the structural and linguistic annotation as possible. Experiments in adapting the tools for linguistic preprocessing (tokenization, part-of-speech tagging, lemmatization) used in the DWDS project to the peculiarities of CMC discourse make up an essential part of the current project work. The final decision about the target size of the corpus will be made as soon as these experiments allow for a reasonable estimation of the amount of time needed for those parts of the annotation, which still will have to be done manually (e.g. a fine-grained subclassification of netspeak phenomena, cf. Section 4).

The integration of DeRiK into the DWDS system may be valuable for various research and application fields and will provide an added value to the user of the resources on the DWDS website. The DWDS search engine allows for searches based on surface forms and on parts of speech (and a combination of both). For each query, results from all (sub-)corpora, which match the query will be presented. The calculation of semantically related terms

(abstracting away from the surface forms) is currently beyond the capabilities of the DWDS search engine. Their encoding is part of a lexicographic description drawing on the various corpora (see below). In particular, we envision the following usage scenarios of the CMC component within the DWDS framework:

(a) Language variation, language change, and stylistics: A general-language corpus that includes a CMC component will provide a broad empirical basis (i) for further corpus-based investigations of the usage and dissemination of CMC-specific phenomena across linguistic varieties and digital genres and (ii) for comparative analyses of the features of CMC discourse and of 'traditional' written genres (e.g. newspaper, fiction, scientific writing, non-literary prose); it will thus facilitate to track and describe how new linguistic patterns and communicative genres emerge.[8]

(b) Lexicology and lexicography: Besides genre-specific discourse markers and 'netspeak' jargon (like 'lol' for 'laughing out loud' or 'imho' for 'in my humble opinion'), new vocabulary is characteristic for CMC discourse, e.g. 'funzen' (an abbreviated variant of the German verb 'funktionieren', en.: 'to function') or 'gruscheln' (verb denoting a function of a German social network platform, most likely a blending of 'grüßen', en.: 'to greet' and 'kuscheln', en.: 'to cuddle'). There are also CMC-specific processes of lexical-semantic changes, e.g. the broadening of the concept of 'Freund' (friend). Up-to-date lexical resources should document and describe these tendencies by integrating CMC data into their data basis. Once the first partial corpora of the DeRiK corpus are made available in the DWDS system, it is intended to extend the DWDS dictionary component with entries describing new lexemes that have evolved from CMC discourse. In addition, the DWDS corpus system will then allow one to track how new vocabulary from CMC discourse (such as the examples aforementioned) spreads into ''traditional'' genres (e.g. newspaper, fiction, nonliterary prose).

(c) Language teaching: CMC has become an important part of everyday communication. Language- and culture-specific properties of CMC should, thus, also be taken into consideration in communicative approaches to Second Language Teaching. In this context, the DeRiK corpus and the documentation of CMC vocabulary in the DWDS dictionary may be useful resources. In school teaching, students with German as a native language may use the DWDS system to compare 'traditional' written language with CMC and to explore how style varies across different genres.

4 Annotation of CMC-Specific Phenomena

One advantage of integrating DeRiK into the DWDS system is that users can profit from the DWDS corpus annotation and querying facilities: The corpus resources that are currently available in the DWDS system are lemmatized with the TAGH morphology (Geyken and Hanneforth, 2006) and tagged with the part-of-speech tagger *moot* (Jurish, 2003). The corpus search engine 'Dialing DWDS Concordancer' supports linguistic queries on several annotation levels (word forms, lemmas, STTS part-of-speech categories) as well as in filtering (e.g. by text type) and sorting options.

As all corpus resources in the DWDS system are encoded according to the guidelines of the Text Encoding Initiative (TEI-P5; TEI Consortium, 2007), the project uses TEI also for the annotation of its CMC component. For this purpose, we have developed a TEI-compliant annotation schema that provides

- a macrostructure of CMC discourse, which covers a broad range of CMC genres (see Section 4.1);
- a partial schema for the description of selected CMC-specific phenomena ('interaction signs': emoticons, interaction words, interaction templates, addressing terms; see Section 4.2 for details).

The discussion in this article will focus on core issues of the schema. A detailed description of the

schema is given in Beißwenger *et al.* (2012).[9] Especially on the microlevel, the schema is open for extensions. For the annotation of the DeRiK data, it will be expanded stepwise with elements for other phenomena, which are described as typical 'netspeak' phenomena in the linguistic literature on CMC. In addition, it is planned to adapt the tools for tokenization, lemmatization, and part-of-speech tagging used in the DWDS project to the peculiarities of written CMC discourse and, thus, enhance the data by automatic annotations of part-of-speech categories and baseforms.[10]

4.1 Annotation of CMC-specific micro- and macrostructures

We introduced the category 'posting' as a basic element to capture CMC micro- and macrostructures. A posting is defined as a content unit that is being sent to the server 'en bloc'. Postings can usually be recognized by their formal structure, even if they have different forms and structures across CMC genres. This facilitates the automatic segmentation and annotation of CMC micro- and macrostructures.

We use the term 'microstructure' to refer to the internal structure of postings. There are cases in which a posting consists of exactly one portion of text. In other CMC genres, e.g. in discussion groups, postings may contain divisions and markup used by the authors to structure their content.

We use the term 'macrostructure' to describe how the postings are sequenced. While microstructures are generated by an individual author, macrostructures do not emerge from the actions of just one user but from all posting activities of all users involved in a CMC conversation plus server routines for ordering the incoming postings.

Our TEI schema distinguishes between two major types of CMC macrostructures:

- 'logfile' structures, which arrange the postings in a linear chronological order based on when they reached the server (as is the case in chats and instant messaging data);
- 'thread' structures, which arrange the postings using two dimensions with specific semantics: the 'above/below' dimension representing a temporal 'before/after' relation; the 'left/right' dimension (by indentation), which usually symbolizes the topical affiliation of one posting to a previous posting (as is the case, e.g., in forum, weblog, and wiki discussions).

4.2 Annotation of 'Interaction Signs'

The corpus-based investigation of 'netspeak' jargon is interesting in many research contexts (style variation and language change, discourse management, language teaching, etc.). Our annotation schema comprises elements for a set of 'netspeak' phenomena, which we term 'interaction signs'. The term builds on the category 'Interaktive Einheiten', which was introduced in the three-volume scientific grammar of the German language Zifonun *et al.* (1997) to classify interjections (such as 'hm' or 'oh my god') and responsives (such as 'yes' and 'no') in spoken discourse. In contrast to part-of-speech-categories, interaction signs are not syntactically integrated and do not contribute to the compositional structure of sentences. They typically serve as devices for conversation management, i.e. they can be used to express reactions to the partners' utterances or to display the speaker's emotions. Besides interjections and responsives, the category 'interaction sign' includes four CMC-specific subcategories (see Fig. 2):

(1) 'Emoticons', which are iconic units that are created with the keyboard and which typically serve as emotion or irony markers or as responsives. Being of iconic origin, the use of emoticons is not restricted to a specific language. However, different styles of emoticons exist—e.g. Western style emoticons such as *:-)*, *:-(*, *;-)*, or *:)*, or Japanese style emoticons such as *(^_^)*, *\(^_^)/*, *(*_*)*.

(2) 'Interaction words', which are symbolic linguistic units whose morphologic construction is based on a word or a phrase. They may describe gestures or facial expressions, e.g. **g** (< 'grins' *grin*), **fg** (< *fat grin*), **s** (< *smile*), or they are used for the simulation of actions and events.

(3) 'Interaction templates', which are units that the user does not generate with the keyboard but which are generated automatically from a file with a previously prepared text or

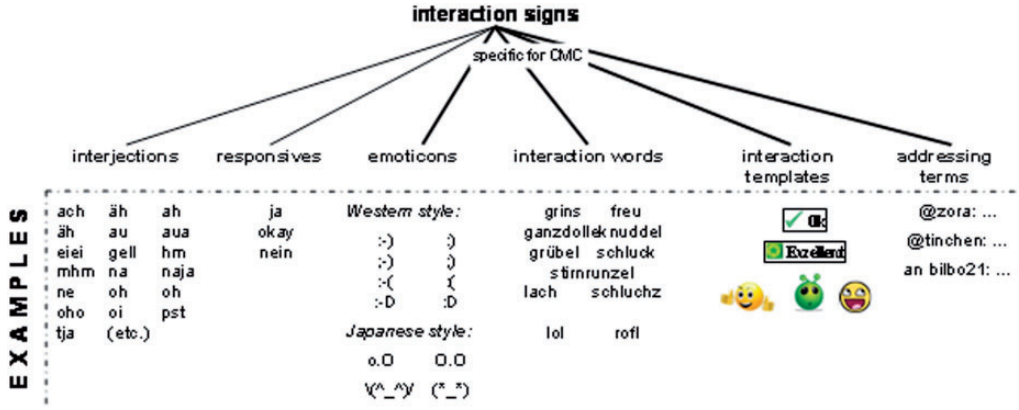

Fig. 2 Typology of interaction signs (with examples)

graphical element after the user has activated a template.

(4) 'Addressing terms', which are units that are used to address an utterance to a particular interlocutor.

For each type of interaction sign, the schema includes a set of attributes, which allow for a fine-grained subclassification of their occurrences in the corpus documents (cf. Beißwenger *et al.*, 2012: section 3.5.1).

5 Conclusion and Outlook

Up to now, many assumptions about the Internet's impact on language change have been based on small data sets. As a new component within the DWDS system, the DeRiK corpus is meant to be a resource for the investigation of language usage in CMC genres on a broader empirical basis. The annotation schema outlined in Section 4 is used and evaluated in the ongoing work of the DeRiK project. The categories proposed in this schema will have to be further discussed within the CMC community. We consider the development of this schema to be a first step towards the development of an annotation standard that will facilitate interoperability between language data and thus cross-language, cross-genre, and micro-diachronic investigations of CMC phenomena on the basis of distributed corpora. The schema focuses on linguistic aspects, but it is open for extensions motivated through other fields of research, i.e. cultural studies or sentiment analysis.

The data that will be collected and annotated in the DeRIK project as well as the tools for their linguistic annotation will be made available through the CLARIN Language Resources and Technology infrastructure.

References

Beißwenger, M. (2013). Das Dortmunder Chat-Korpus. *Zeitschrift für germanistische Linguistik*, **41**(1): 161–64.

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

Beißwenger, M., Ermakova, M., Geyken, A., Lemnitzer, L., and Storrer, A. (2012). A TEI schema for the representation of the computer-mediated communication. *Journal of the Text Encoding Initiative*, **3.** http://jtei.revues.org/476 (doi: 10.4000/jtei.476) (accessed 30 April 2013).

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

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

Forsyth, E. N. and Martell, C. H. (2007). Lexical and Discourse Analysis of Online Chat Dialog. *Proceedings*

of the First IEEE International Conference on Semantic Computing (ICSC 2007). Irvine, California: IEEE Computer Society, pp. 19–26.

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

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

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

Herring, S. C. (ed.) (2010). *Language@Internet*, **7.** http://www.languageatinternet.org/articles/2010 (accessed 30 April 2013).

Jurish, B. (2012). *A Hybrid Approach to Part-of-Speech Tagging. Final report, project 'Kollokationen im Wörterbuch'*. Berlin: BBAW. http://www.dwds.de/dokumentation/tagger/ (accessed 28 June 2013).

Kestemont, M., Peersman, C., De Decker, B., De Pauw, G., Luyckx, K., Morante, R., Vaassen, F., van de Loo, J., and Daelmans, W. (2012). The Netlog Corpus. A resource for the study of flemish dutch internet language. *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Paris, pp. 1569–72.

King, B. W. (2009). Building and analysing corpora of computer-mediated communication. In Baker, P. (ed.), *Contemporary Corpus Linguistics*. London: Continuum, pp. 301–20.

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

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

Reynaert, N., Oostdijk, O., De Clercq, H., van den Heuvel, H., and de Jong, F. (2010). Balancing SoNaR: IPR versus Processing Issues in a 500-Million-Word Written Dutch Reference Corpus. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. Paris, pp. 2693–98.

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

Storrer, A. (2013). Sprachstil und Sprachvariation in sozialen Netzwerken. In Frank-Job, B., Mehler, A., and Sutter, T. (eds), *Die Dynamik sozialer und sprachlicher Netzwerke. Konzepte, Methoden und empirische Untersuchungen an Beispielen des WWW*. Wiesbaden: VS Verlag für Sozialwissenschaften, pp. 329–64.

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

Zifonun, G., Hoffmann, L., Strecker, B., Ballweg, J., and Brausse, U. (eds), (1997). *Grammatik der deutschen Sprache*, **3.** Bände. Berlin; New York: de Gruyter.

Notes

1 http://www.empirikom.net/bin/view/Themen/DeRiK.
2 http://www.empirikom.net/.
3 http://www.chatkorpus.-tu-dortmund.de.
4 http://faculty.nps.edu/cmartell/NPSChat.htm.
5 https://wiki.itmc.tu-dortmund.de/cmc/.
6 https://groupes.renater.fr/wiki/corpus-ecrits-nouvcom/index.
7 This dictionary is based on a six-volume paper dictionary, the 'Wörterbuch der deutschen Gegenwartssprache' (WDG, en.: 'Dictionary of Contemporary German'), published between 1962 and 1977 and compiled at the Deutsche Akademie der Wissenschaften (Klappenbach and Steinitz (eds.) 1964-1977).
8 Overviews of the features of CMC discourse from a linguistic perspective can be found, e.g., in Herring (ed., 1996, 2010), Runkehl *et al.* (1998), Crystal (2001, 2011), Beißwenger and Storrer (2008), and Storrer (2013).
9 The RNG schema file, a TEI-compliant ODD documentation as well as encoding examples are available at http://www.empirikom.net/bin/view/Themen/CmcTEI.
10 For this task, the project cooperates with the scientific network 'Empirical Research on Internet-Based Communication' (Empirikom, http://www.empirikom.net) funded by the Deutsche Forschungsgemeinschaft (DFG) as well as with the research project 'Corpus-based linguistic retrieval and analysis with the help of Data Mining' (*KobRA*, http://www.kobra-tu-dortmund.de) funded by the German Federal Ministry of Education and Research (BMBF).