Preprint. To appear in: Anke Lüdeling & Merja Kytö (eds):
Corpus Linguistics. An International Handbook.
Series: Handbücher zur Sprach- und Kommunikationswissenschaft/
Handbooks of Linguistics and Communication Science. Mouton de Gruyter, Berlin.

21. Corpora of Computer-Mediated Communication

Michael Beißwenger & Angelika Storrer

1. Introduction

„Computer-Mediated Communication (CMC)" is a research field that explores the social, communicative and linguistic impact of communication technologies, which have continually evolved in connection with the use of computer networks (esp. the Internet). Early work on CMC investigated the use of e-mail, mailing lists, Usenet discussion groups, Multi User Dimensions/Dungeons (MUDs) and Internet Relay Chats (IRC) (e.g. Reid 1991, Maynor 1994, Vilmi 1994, Herring 1996). Since the mid-90ies, when web-based CMC tools made computer-based communication even easier to use and more widespread, the investigation of features and processes particular to CMC has established itself as an independent research field with specialized journals and workshops, e.g. *Journal of Computer-*

Mediated Communication, *Journal of Interactive media in Education*, *Journal of Online Behaviour* and *Language@Internet* or the minitrack on persistent conversation at the annual *Hawaii International Conference on System Sciences* (*HICSS*). By browsing John December's portal of CMC resources and research (http://www.december.com/cmc/info/), one may get an impression of the broad range of CMC-related topics, as well as of the rapid change and the continuous evolution of new CMC tools and genres, such as Instant Messaging Services (e.g. *ICQ*; Leung 2001, Grinter/Palen 2002, Leung 2002), weblogs (e.g. Herring/Scheidt/Bonus et al. 2004) or Wiki-based applications such as „wikipedia" (e.g. Viegas/Wattenberg/Dave 2004).

CMC research is an interdisciplinary field: psychological issues such as motivation, understanding and decision-making in CMC environments are investigated within the context of e-learning and professional collaboration via the Internet. Social science categories such as identity, gender, trust, and confidence are crucial for studies of online communities and social networks. Linguists investigate how language is used in computer-mediated settings and how the technical and pragmatic conditions of the underlying technology affect the strategies of language production and understanding (Herring 2001, Herring 2004). Crystal (2001) uses the term „netspeak" to cover the broad range of peculiarities and specific features of discourse produced on the Internet, the most well-known of them being emoticons, acronyms (*IMHO* = *in my humble opinion*, *AFK* = *away from keyboard*), speedwriting and abbreviations (*4* = *for*, *g* = *grin*) and conventions for simulating prosodic features (e.g. upper case = loud voice). Research on discourse structure in CMC usually distinguishes between two main genres, namely synchronous and asynchronous CMC (cf. Herring 2001, 614 p.). In synchronous CMC, such as chat or Instant Messaging, people exchange messages instantaneously and in real-time; all participants are simultaneously online and react immediately and only with a slight delay to messages from other participants. In asynchronous CMC, such as e-mail, mailing lists or discussion groups, users do not have to be online at the same time to communicate; the addressee of a message may both read and respond to it at a later time.

A central issue in the linguistic CMC-related discussion is the status of computer-mediated discourse relative to the distinction of orality and literacy, especially its status within the dichotomy of writing and speaking. This has to do with the observation that written language produced by means of CMC devices, although represented graphically, often shows features particular for speech and interactive discourse and, thus, differs stylistically from „traditional writing" in similar genres. On the other hand, synchronous text-based CMC, as found in chat environments, differs considerably from face-to-face conversations (Crystal 2001, Storrer 2001, Zitzen/Stein 2004). Although methods and categories developed in conversation analysis play a central role in describing how people manage to establish and maintain interactional coherence in CMC discourse, obviously central notions – such as „turn", „turn-taking", „speakerhood" and „floor-holding" – have to be adapted and reconsidered, especially with regard to synchronous CMC (Herring 1999, Beißwenger 2003, Zitzen/Stein 2004).

The focus of our article is on corpora that may serve to support empirical research on linguistic aspects of CMC discourse. In section 2 we present a short overview of different types of existing CMC corpus resources and give examples of how they have been used to

study „netspeak" properties and other CMC peculiarities. In section 3 we discuss problems and issues with which designers of CMC corpora are confronted, especially when building resources that are linguistically annotated and enhanced with metadata.

The main focus of current CMC research lies on Internet-based technologies and their genres (including videoconferencing and multimedia applications). However, a clear line between these genres, genres of telecommunication (such as SMS) and other more „traditional" media (interactive television, Internet broadcasting) can not be drawn, because computer technology and other communication and entertainment media coalesce in the digital age. In this article we concentrate on corpora of prototypical Internet-based discourse genres such as e-mail, mailing lists, chats and discussion groups, which reveal typical „netspeak" peculiarities and have been collected in the context of CMC research.

The corpora mentioned in our overview are not intended to form a complete list; the selection was motivated by the goal of finding prototypical examples for our typology. The selection given below is accompanied by a list of WWW resources available at http://www.computervermittelte-kommunikation.de/020_ressourcen/030_corpora.

2. Overview of CMC corpora

For our overview we found it useful to establish some classification criteria that serve as the basis for a typology of existing CMC corpus resources. (For the interdependence between design strategies and corpus types cf. article 11)

Initially, one may differentiate between *project-related corpora* and *corpora for general use*. The former were compiled as an empirical basis for questions in a particular project; the latter do not directly pertain to a particular project, but were established rather as a data pool for the investigation of diverse potential research questions. Depending on the amount of work spent structuring and annotating the corpus data with respect to potential research questions, one may, furthermore, differentiate between *corpora of raw data* and *annotated corpora*. In *annotated corpora*, the data have been subjected to annotation processes (e.g. an SGML/XML-based annotation of data segments that may be relevant for purposes of analysis) (For issues in corpus pre-processing cf. articles 24-34). In contrast, the data in *corpora of raw data* have been left in the condition in which they were originally acquired from the Internet. Figure 1 shows the four main types of CMC corpora, which result from these classification criteria; in the following we will illustrate each type using prototypical examples.

Data edited for purposes of analysis? / The corpus was originally designed to be …	**No**	**Yes**
... project-related	**1** ***corpora of raw data***	**3** ***annotated corpora***
... for general use	**2** ***corpora of raw data***	**4** ***annotated corpora***

Figure 1: types of CMC corpora.

2.1. Type 1: project-related corpora of raw data

Research on Computer-Mediated Communication is presently conducted for the most part with *project-related corpora of raw data*. This is due to the fact that CMC research is a relatively new field and that CMC genres currently are not at all or only marginally represented in large balanced corpora. Thus, at the present time, the assortment of large accessible corpora that were exclusively designed for analyzing CMC phenomena is rather unsatisfactory. Therefore, for empirical studies, corpora often have to be individually acquired from the Internet or obtained from users of CMC facilities. Examples of such type 1 corpora are, for instance, Yates' *CoSy:50*-Corpus, which comprises 50 submissions each, from 152 computer conferences (described in Yates 1996), Todla's *Thai Chat Corpus* (used in Todla 1999 and Panyametheekul/Herring 2003), Janich's corpus with messages from a university mailing list (used in Janich 2002), Berjaoui's corpus with logfiles from Moroccan IRC channels (cf. Berjaoui 2001), the extensive Swiss-German webchat corpus (15,86 million words) recorded by Siebenhaar (cf. e.g. Siebenhaar 2006) or Pankow's contrastive German-Swedish *IRC-Corpus* (used in Pankow 2003). Type 1 corpora usually have a manageable size and are often not compiled with a potential third party user in mind (who possibly pursues other research questions). Accordingly, corpora of that kind are often just rudimentarily documented (i.e. only in the publications in which they are cited). Moreover, they are usually only accessible through the scholars who compiled them for their own purposes.

2.2. Type 2: corpora of raw data for general use

Contrary to their project-related equivalents, *corpora of raw data for general use* are compiled in order to give scholars a data pool for the empirical study of diverse research questions. Examples are the *Netscan Usenet database* (http://netscan.research.microsoft.com/), the *Korpus deutschsprachiger Newsgroups* (cf. Feldweg et al. 1995), the *Enron Email Dataset*, which holds over half a million business-related e-mail messages (http://www-2.cs.cmu.edu/~enron/), the *WWE-2006 weblog dataset* which was temporarily available to the participants of the *3rd Annual Workshop on the Weblogging Ecosystem* (see http://www.blogpulse.com/www2006-workshop/), and the *SpamAssassin Public Corpus* (http://spamassassin.apache.org/publiccorpus/) from the *Apache SpamAssassin Project*, in which approximately 6,000 e-mail messages were compiled as a discretionary data pool for research in automatic e-mail classification.

2.3. Type 3: project-related annotated corpora

Project-related annotated corpora are compiled with CMC data in order to empirically conduct a particular research project. When building up corpora of this type, the data are subjected to a coding process, which facilitates both the work with the corpus and the access to and analysis of the data. An example of a type 3 corpus is the e-mail corpus from Declerck and Klein (cf. Declerck/Klein 1997) collected in the framework of the *COSMA*

project and comprising 160 e-mail messages with appointment arrangements. The data of this corpus have been syntactically and semantically annotated with the research objective in mind – the development and evaluation of the analysis component of a dialogue system for appointment scheduling. Another type 3 example is e.g. Paolillo's 24-hour logfile from an IRC channel, in which types of messages and selected linguistic features were coded in order to apply quantitative methods for the analysis of textual CMC data (cf. Paolillo 1999). In Naumann's corpus (used e.g. in Naumann 2005), logfiles from „virtual classrooms" have been linguistically annotated for an analysis of conversational rules in educational chat applications. A large type 3 corpus of personal homepages has been compiled in Rehm's *Hypnotic* project in order to develop procedures for automatic web genre classification. Aside from HTML documents, *Hypnotic* also contains files with news articles, as well as over 1,000 e-mails archived from the WWW (cf. Rehm 2002, 2006). Another large annotated corpus of websites (a „web genre document bank") has been compiled within the *Indogram* project. The particular design and annotation principles are described in Mehler/Gleim (2005).

2.4. Type 4: annotated corpora for general use

At present, there are few corpora whose data pool has been collected *ab ovo* as a resource for CMC research and has been edited with regard to various possible research questions. The *Düsseldorf CMC Corpus* (as described in Zitzen 2004, 14-36) contains grammatically annotated data from various synchronous and asynchronous CMC genres (e-mail, mailing lists, newsgroups, guest books, chats) and has a range of approximately 230,000 tokens. The *Dortmund Chat Corpus* (http://www.chatkorpus.uni-dortmund.de) comprises approximately 450 logfiles with a total of some 990,000 tokens from various chat applications (chats in professional contexts such as, for example, teaching, learning, counseling and media contexts, as well as „social chats" within and outside of media contexts) and has been processed for linguistic research purposes on the basis of an XML annotation.

3. Collection and linguistic annotation of CMC-corpora

The mere acquisition of data for CMC corpora can be accomplished rather easily. With little effort, linguistic data can be saved from public access CMC platforms or from the archives of client programs of e-mail, news and instant messaging services, IRC- or MUD-facilities, which have been made accessible to the researcher. The sheer availability of data that can easily be saved does not imply that the design of CMC corpora for linguistic research purposes is trivial. Instead, the design of CMC corpora involves an array of challenges and preliminary decisions that do not arise in a similar way when designing corpora of „traditional" text or speech genres. This is due to, on the one hand, the communicative frameworks of CMC genres and, on the other hand, the embedding of CMC data in Internet-based storage and presentation formats.

3.1. Challenges in data acquisition and documentation

Above all, questions relating to the management and structuring of data are crucial to the acquisition and documentation of content for CMC corpora. Furthermore, with regards to purposes of analysis, the question arises as to which technical, conversational and sociological meta-information should be captured during the collection of the data and which ethical aspects should be considered when designing CMC corpora.

3.1.1. Data sampling

Indeed, textual CMC data are available in large numbers and are simple to archive; however, the data for a CMC corpus should be acquired dependent on the purposes that the corpus should serve (for project-related corpora, e.g., with regards to the pursued research questions). Should the data be compiled by the principle of convenience, randomly or according to particular linguistic phenomena? Herring (2004) gives a detailed overview of the advantages and disadvantages of various data compilation techniques for the purpose of Computer-Mediated Discourse Analysis. (For issues involved in data collection in general cf. article 11)

3.1.2. Format of the original data

Language data found in individual CMC environments can either be saved in their original form (i.e. including all possibly relevant layout and structure information) or in reduced form (reduced to pure character strings). The decision as to which starting format one would like to use in corpus design should depend on the intended or potential research questions, for the processing of which the corpus should deliver the empirical basis. Especially with regards to CMC genres that use the layout and multimedia properties of WWW-based online publishing, it may be relevant to capture the data not only in the form of pure character strings, but rather as a kind of rich data, which also include layout information such as typeface, font, color, size, manual vs. forced line break, paragraph structure, etc. However, layout properties of written contributions in WWW-based discussion forums, chats and in newer e-mail client programs can adopt pragmatic functions and, thus, are often more than just text decoration within communicative exchanges. Therefore, when choosing a reduced starting format, one should carefully take into consideration that the archived data may have lost some visual properties that were functional in the original context. Furthermore, since the use of graphic elements within CMC environments has become more and more common over the last years, it may be relevant to also represent certain graphic encoding formats in the corpus. In online forums, guest books, weblogs and webchats, graphics can be used as a substitute for emoticons, e.g. by selecting them per mouse click from a selection menu. These graphics – like emoticons – can assume evaluative, expressive or regulative functions (cp. Runkehl et al. 1998, 98 p.), sometimes even illustrative or emblematic functions. If data for CMC corpora are captured in ASCII-format by simply saving screen content, then not only potentially relevant layout

information becomes consequently lost, but graphics also do not remain preserved. Incidentally, this is not merely applicable to WWW-based CMC applications, but also to e-mail and other non-WWW-based CMC applications, whose client software allows the integration of media objects (graphic, audio, video).

3.1.3. Representation format

Depending on the starting format, it is important to choose an appropriate representation format for the language data contained in the corpus. If the original data were generated with diverse CMC systems (e.g. on the basis of different forum or chat software) and/or represent different CMC genres, then one should choose a representation format that accommodates all of these systems and genres. With regards to a long-term data storage, one should choose an interchange format that can be processed and converted regardless of specific applications or special software (e.g. a description format based on an SGML or XML language). Furthermore, a decision must be made as to whether and in which way meta-information that is already contained in the original data should be transferred into the chosen representation format. This applies e.g. to the so-called „header information" in e-mails (cf. Evert/Fitschen 2001, 371 p.). In this context, the establishment of a standardized framework for the representation and exchange of CMC data is a strong desideratum (cf. Bruckman et al. 2000).

3.1.4. Capturing hard-coded references (e.g. hyperlinks)

In the acquisition of data and their representation in the corpus, non-sequential structures, if applicable, must also be taken into consideration: communicative contributions that are captured from WWW-based CMC environments may contain hyperlinks, with which reference is made to external resources. The reconstruction of the content of these resources may be important for the corpus-based analyses of the respective contributions. Furthermore, the structural embedding of a contribution within its context (e.g. of a single posting within the thread structure of a forum) can, in its original presentation format (e.g. the WWW-interface of the respective forum system), also be organized through hyperlinks. On the one hand, there are *system-generated hyperlinks*, which embed the single contribution into the structural and/or thematic organization of the CMC application. On the other hand, there are *hyperlinks that are manually defined* by the users. System-generated hyperlinks would be e.g. those that link single postings in an online forum together into threads and, thus, create a navigation design, which makes all contributions to a thread accessible per mouse click. Likewise, hyperlinks that (a) mark a posting as a reply to a preceding posting by another author and (b) make the latter accessible via mouse click can also be deemed system-generated. An example of a manually set hyperlink would be e.g. a hyperlink that is inserted by the author in the text of his posting in order to refer to a posting in another thread of the forum or to an external website. If such referential structures are not represented in the corpus, then coherence relations that refer beyond the single communicative contribution cannot be reconstructed.

Moreover, when saving CMC data, one should keep in mind that hyperlinks on the user interface are not always simply displayed in the form of a clickable target resource URL, but may also appear in the form of textual link buttons (e.g. „for more information, click here"). In such cases, the URL information can only be obtained from the source code. When saving just the screen data, the respective URL information becomes lost.

3.1.5. Capturing implicit references (e.g. cross media)

In some cases, CMC facilities in the media context can correspond with and accompany media events outside of the Internet. Examples would be chat events accompanying the matches in a soccer championship or the Tour de France, or moderated chat events with studio guests from a preceding TV show (specifically, talk shows with political or „infotainment" formats). In these types of CMC events, topics that were initiated outside (e.g. in the TV broadcast) may be perpetuated or taken up again within the particular CMC environment. Furthermore, certain issues addressed within the CMC event may only fully be understood by examining the corresponding passages of the parallel TV broadcast. In teaching and learning contexts, CMC facilities are sometimes supplemented through resources that are stored on external platforms – e.g. on an accompanying website or in a „materials" section on an e-learning platform. Users of the particular CMC environments may not explicitly link to these external resources when they discuss them because they can assume that each user knows where the respective materials are located. It may be important to either incorporate the respective resources (each as a whole) into the corpus or to at least have precise descriptions of them available; while doing corpus analyses, this would help resolve linguistic references made to them. The same goes for chat conversations that are conducted using a so-called „shared whiteboard system". A „shared whiteboard" is a kind of virtual chalkboard, on which – parallel to the chat activity – graphics, visual aids and tasks can be provided for the participants, drawings can be created or documents can be collaboratively edited. The content of the whiteboard is not usually documented in the logfile. Therefore, when capturing data from such systems, it may be necessary to decide to which extent the content displayed or produced on the whiteboard (or a description of it) should also be included in the corpus.

3.1.6. Capturing metadata concerning the communication environment

Technical factors that are formative for the structure of communicative events must be taken into consideration when documenting corpus data (e.g. program functions of the respective chat or forum system). Without this information, certain linguistic and interactional characteristics of the data often cannot be interpreted adequately. The same holds for socio-communicative roles (e.g. the role of a forum or chat moderator) and the communicative authority connected with these roles (e.g. the authority to delete or edit messages written by other users, the authority to revoke a user's right to post further messages), which either apply to an entire CMC platform or are only defined for a singular communicative event on the respective platform (e.g. within the framework of a single

expert chat conducted on the e-learning platform of a blended learning seminar). Depending on the technical conditions, the linguistic form and the conversational structures in one and the same CMC genre can sometimes vary considerably, as can be seen in, for example, a comparison between non-moderated chats (e.g. social chats on IRC or the web) and technically moderated chats (e.g. celebrity chat events). The contributions in both of these subtypes of chat discourse differ not only with respect to their average length (5.38 versus 17.22 tokens per message, according to an unpublished spot survey by the authors); they also differ with respect to the frequency of split moves: the partitioning of a communicative move onto several messages does not occur in technically moderated chats (cf. Beißwenger 2003, 219-224). Moreover, and also due to the technical framework, the contributions not only tend to be much longer in the technically moderated chats, but also much more elaborate than contributions in other types of chat applications. The patterns of communicative moves that can be observed in technically moderated chats are also much more rigid than those in non-moderated chats.

3.1.7. Capturing sociological meta-information

Sociological meta-information about the CMC user groups is of great importance for the field of sociolinguistic and socio-psychological CMC research. The analysis of corpus data and the differentiation between group-specific stylistic features and ranges of linguistic variation (cf. Androutsopoulos 2003) will strongly benefit from information about the users of the CMC environments from which the corpus data were recorded. This includes e.g. information about the users' average age, gender and level of education, about the degree of acquaintance among the users and the degree of attachment to the respective platform (as a member of an „online community"). A typological approach to „online communities", which combines the technical and communicative frameworks, as well as pragmatic and sociological meta-information about CMC platforms, can be found in Porter (2004).

3.1.8. Questions concerning research ethics

Ethical questions should be taken into consideration in the creation of CMC corpora in order to protect the personal rights of the users recorded in the corpus data. In order for the research to remain ethically justifiable, it would be ideal to inform people when they or their statements are being recorded for research purposes. Due to reasons of practicability, when collecting data from publicly accessible CMC environments, it is unrealistic to obtain a declaration of consent for the recording and subsequent use of users' statements for research purposes, since such environments often have many, and sometimes even frequently alternating, users. Moreover, an *informed consent,* received in advance, would in many cases compromise the authenticity of the communicative behavior of the users. However, obtaining the consent of the participants retrospectively is often just as unrealistic, since the participants are registered under pseudonyms. Although users of publicly accessible IRC, MUD, webchat and forum environments already operate under

these self-chosen pseudonyms („nicknames"), that does not necessarily mean that data recorded in such environments may be used without anonymizing them for research purposes. After all, a third party may happen to meet the respective users when logging onto the environment from which the corpus data has been recorded. Furthermore, in CMC environments in which the users can create „user profiles" with personal information, there is even a possibility that the real person behind the CMC character could be identified through a targeted inspection of the contact information provided by the user (e.g. e-mail address, ICQ or telephone number). Whether a retrospective anonymization of the participant names or a mere omission of details about the particular CMC platform's location is more appropriate from an ethical point of view – this question is still being discussed and approached differently by various scholars: „some feel that they have a moral obligation to obtain explicit permission from the authors for *publishing* logs in academic papers (…); others collect logs without asking for permission but the logs are then only processed by statistical software and not read by humans (…); many others simply do not declare explicitly whether permission was obtained for their logs" (Paccagnella 1997). With regards to research ethics, Paccagnella stresses even further, „changing not only real names, but also aliases or pseudonyms (where used) proves the respect of the researchers for the *social reality* of cyberspace." A good overview of the discussion about how to handle CMC data in an ethically justifiable manner – in the frame of which questions concerning authorship and copyright are also addressed – is provided e.g. by Bruckman et al. (2000), Crystal (2001, 191-194) and Döring (2003, 236-242).

3.2. Challenges in data editing and annotation

When annotating CMC data, one should both carefully develop appropriate description categories and document grammars, which grasp the linguistic particularities of CMC genres, and modify existent tools for the linguistic preprocessing of speech data (e.g. morphosyntactic taggers).

3.2.1. Description categories

Categories that were developed for the annotation of structural and functional units in „traditional" discourse genres (text and speech genres) cannot be used for the description of CMC discourse without first being adapted with regard to the particularities of CMC. Thus, there is a need for categories that account for the unclear position of CMC between orality and literacy. Thereby, central concepts of the discourse analysis must be evaluated and reinterpreted. For instance, in synchronous CMC, simultaneous backchannel feedback is not possible – but this does not inevitably mean that in synchronous CMC there are absolutely no functional *equivalents* to the backchannel behavior in face-to-face conversations. Likewise, one should consider to which extent it is appropriate to describe conversation structures in synchronous CMC by uncritically using categories such as „turn", „turn taking" and „sequentiality" (cf. e.g. Murray 1989, Werry 1996, Garcia/Jacobs 1999, Herring 2001, Beißwenger 2003).

3.2.2. Interpretative description of conversation structures

When annotating conversation structures in logfiles from synchronous CMC genres, one should keep in mind that the reconstructive assignment of the participant's contributions to thematic threads or patterns of communicative moves (e.g. „question-answer") is always a matter of the researcher's individual interpretation. Due to the lack of para- and nonverbal data and due to the technical (not pragmatic) sequencing of participant submissions (see section 3.3), there is a larger margin for interpretation (or speculation) in the modelling of CMC data than in the modelling of data from (oral / face-to-face) conversations. Example (1a) shows a sequence of two messages from a one-to-one chat in the context of psychosocial counselling. The respective logfile was kindly provided by the project „Psychosoziale Hilfe online" (cf. van Eckert 2005).

Due to its move types („question" and „assertion"), the sequence could be interpreted as the realization of a „question-answer" pattern. But when analyzing this example in its larger context, as provided in (1b), the before-assumed answer („my mother is going crazy") actually is revealed to be just part of an answer to an earlier question, which was broken up into several messages by the respective user. Due to the lack of means for simultaneous coordination among the participants, it is displayed after the communication partner's next question. This interpretation of example (1b) is substantiated when one considers the message's time of arrival at the server, which is documented by some chat systems in the form of a so-called „timestamp" in the logfile. It is quite unlikely that B could have received, read and answered the question from A „may I ask how old you are?" within 2 seconds.

(1a)

***A*:** darf ich fragen, wie alt du bist?
may I ask how old you are?
***B*:** meine mutter is hysterisch
my mother is going crazy

(1b)

Timestamp	*message*
15:52:42	***A*:** wie sieht denn die krise in der familie bei dir aus? *tell me about the crisis in your family.*
15:53:08	***B*:** meine ma redet nicht mit meinemdad und andersrum *my mom won't speak to my dad and vice versa*
15:53:22	***B*:** mein bruder redet nicht mit ihnen und andrsrum *my brother won't speak to them and vice versa*
15:53:37	***B*:** ich rede wenig mit ihnen und sie gar nicht *i hardly speak to them and they don't speak at all with me*
15:53:50	***B*:** meine eltern wollen sich scheiden *my parents want to divorce*

15:54:01	***A*:** darf ich fragen, wie alt du bist? *may I ask how old you are?*
15:54:03	***B*:** meine mutter is hysterisch *my mother is going crazy*
15:54:06	***B*:** 12 *12*

3.2.3. Linguistic preprocessing and annotation

Tools developed for the automatic annotation of linguistic data (sentencizers, POS taggers, lemmatizers, chunk parsers) cannot be used for processing CMC data without being adapted. This problem is due to some characteristics of language use in CMC that we briefly mentioned in our introduction: speedwriting (e.g. *you* > *U*, *two/too* > *2*, *please* > *plz*; cf. Danet 1997), non-standard spellings (e.g. Engl. *out of* > *outta*, *see you* > *cee ya*; cf. Crystal 2001, 164 – or French *quelqu'un* > *qqn*, *c'est* > *c*; cf. Werry 1996, 55), highly colloquial (slang) or conceptionally oral forms (in German e.g. *ne* (< *nicht wahr?*), *haste* (< *hast du*), *willste* (< *willst du*), in English e.g. *gonna*, *gotta*), letter repetition as a means of emulating prosody (*uiiiiiii*, *sooooo*, *hellooooooo*), written dialect (which may be used either non-intentionally or to mimic the style of particular discourse communities, e.g. of Australian speakers of English; cf. Werry 1996, 58) and abbreviations (*btw* for *by the way*, *lol* for *laughing out loud*, *aka* for *also known as*). In order to successfully process CMC data, the search patterns or lexica of the respective tools must be extended to include these typical „netspeak" elements and treat them appropiately.

Furthermore, the frequent omission of upper case all may lead to false categorizations when using tools that were developed for parsing „traditional" text genres (cf. Ooi 2001). This is especially important when processing languages in which the distinction between lower case and upper case signifies part-of-speech information, as e.g. in German orthography, where nouns are capitalized. Moreover, the creative or expressive use of punctuation marks and special characters must be taken into account: emoticons emulate typed facial expressions by means of punctuation marks and special characters; verbalizations of gesture or physical action (engl. **smile**, **grin**, **hugs**, **shakes hand**; germ. **knuddel** "hug", **tassekaffeeanbiet** "offercupofcoffee") are often enclosed in asterisks (cf. e.g. Reid 1991, Werry 1996, 60, Runkehl et al. 1998, 106 p., Beißwenger 2000, 105-116). Likewise, typography and layout can provide significant characteristics for identifying typical CMC categories (e.g. angle brackets or bold type and a preceding forced line break to distinguish strings of the type „nickname"). Unusual word order and elliptic constructions in CMC discourse may, in addition, pose problems for frequency-based part-of-speech taggers that have been trained on corpora with written standard language, e.g. newspaper corpora (see article 25).

All in all, tools for corpus search and annotation, developed with standard orthography and „traditional" text genres in mind, need to be adapted to the peculiarities of CMC discourse genres.

3.3. The status of logfiles of synchronous CMC and their value as source material

The starting point and empirical foundation of CMC research are those data that are exchanged between users of CMC systems for the purpose of communication and that can be saved through one of the users' computers or through the intermediary server. In synchronous text-based CMC, these types of data are usually organized and displayed in the form of scroll-like records (the so-called *logfiles*). The way in which the individual users' messages are arranged in these records can vary; in standard systems (as used in IRC or many social chats on the web), messages are arranged by a „first come, first served" principle, according to the chronological order in which they reach the server. This generates coherence problems such as disrupted adjacency (Herring 1999) or scrambled threads (Storrer 2001a), which are crucial for synchronous CMC discourse and which, therefore, must be taken into consideration for the analysis.

When acquiring logfiles of synchronous CMC for the design of CMC corpora, one must consider whether one is dealing with *server logs* or *client logs*. The former are recorded directly from the server and represent the complete communicative activity of a „channel", „room" or „separée" of the respective CMC platform; the latter can be created by capturing the communicative activity, which can be observed on one of the users' computers and which, in many cases, only represents the user's individual view of the communicative event in which he is involved. Most of the synchronous CMC systems operate with the use of the social deictic *you* in the automatically generated system messages, which display the users' changes in their communicative status (e.g. *You entered the channel Detroit Rock City*). Therefore, when using client logs, it is advisable to document whose view of the communicative event is being represented in the logfile.

Furthermore, one must consider to what extent data, which can be acquired from the users' screen or the intermediary server and which can be fixed through saving, are sufficient when the purpose is to investigate how synchronous CMC users organize their exchange. The term „synchronous CMC" means that the users must be logged in synchronously on the particular CMC platform; it does *not* mean that communication in webchat, IRC or Instant Messaging proceeds *simultaneously*. Rather, messages are first produced, then subsequently sent to the server and, lastly, transferred en bloc from there to the addressees' computers. Information about when, for how long and, if applicable, with which disruptions and revisions a message is produced is completely indiscernible for the other participants. A message first becomes visible when it appears on the screen as a text block. Therefore, in synchronous CMC – contrary to face-to-face conversations – a runtime negotiation of turns is not possible. Likewise, which messages have already been read by a user during the production of his/her new message (and which have not) is not documented in the logfile at all. Merely the fact that something is visible on the screen does not necessarily mean that the user in front of the screen has taken notice. The difference to face-to-face conversation becomes clear here as well: while it is virtually impossible not to perceive something that is orally expressed by a communicative partner, it could very well happen that something displayed graphically on a screen is not at all or not directly perceived by its addressee (e.g. because he is composing his own message and, hence, looking at the keyboard or the text submission field and not at the logfile) (cf. Beißwenger

2003).

Thus, in many regards, logfiles of synchronous CMC render considerably less information about the conversational process than transcripts of verbal conversations. While the latter are prepared by linguists through an intellectual interpretation of audio or video recordings for the purpose of linguistic analysis, logfiles at best have the status of machine-made „recordings". They neither represent the communicative moves in the sequence in which they were intended by the participants, nor do they document the process of message production – instead, messages in logfiles are represented as text blocks that do not analogously develop over time. Features on the linguistic surface of messages and rhetorical relations between them can, thus, only be explained through speculation, as long as the analyst is left to rely solely on the data contained in logfiles. Therefore, for various purposes of analysis, it is wise to collect additional data aside from just logfile recordings. In order to gather data about the message production processes, Garcia & Jacobs (1999) base their analysis of discourse organization in IRC on video recordings of the users' screens. Ogura and Nishimoto (2004) or Vronay, Smith & Drucker (1999) incorporate typing histories and client-side logging of keytrokes and mouse actions as additional data in their investigation of synchronous text-based CMC. Jones (2001) analyzes series of screenshots from the users' desktops („screen movies"), in order to investigate how users manage multiple and parallel conversations. A research design for the collection of logfile data in combination with video screen capturing and a video observation of the users is described in Beißwenger (2007).

4. Conclusion and future perspectives

With the increasing amount of reading and writing that people do on the Internet, corpus designers who set out to provide balanced corpora that include all relevant text types of contemporary language, should henceforth include CMC discourse as well. At present, online text corpora of contemporary language, such as the *British National Corpus* (*BNC*, http://www.natcorp.ox.ac.uk/), do not; neither do the German corpora available through the *COSMAS* search tool (http://www.ids-mannheim.de/cosmas2/) or the online corpus of the German language of the 20th century compiled within the framework of the online dictionary project *DWDS* (http://www.dwds.de). In „older" corpus collections, such as the *Brown Corpus* (available through ICAME, http://nora.hd.uib.no/whatis.html), the reason for the lack of CMC genres is simply that at the time of the corpus creation CMC was non-existent. In more recently collected language corpora, such as the above-mentioned, the disregard for CMC in corpus building may be due to the unclear status of CMC discourse with regard to the spoken–written dichotomy. Since this dichotomy is crucial for the categorization in speech and text corpora, it is difficult to decide whether CMC discourse should form part of text or speech corpora.

It may have become evident by our review, that the compilation and annotation of CMC corpora may be regarded as a challenging task, for which special methods and devices are needed. Standard tools for linguistic annotation of text corpora are only suitable to a limited extent for CMC corpora, since in interactive CMC discourse punctuation marks are often used in a non-standard way or are even completely neglected. Typing errors and

typical „netspeak“ items (abbreviations, smileys and the like) create obstacles for morphological analyzers or lemmatization tools. Furthermore, the peculiarities of CMC discourse compromise the precision and recall of corpus search tools. Merely the automatic recognition of sentence boundaries requires robust techniques that accommodate for the informal use of punctuation marks typical for CMC. System-generated markup, e.g. the labeling of quotations in e-mails and forum postings and/or the marking of nicknames and system messages in chats, have to be filtered out and separated from the actual CMC discourse. With regard to the ethical aspects addressed in section 3.1.8, the archiving of CMC discourse should preferably be carried out with the permission of the participants. If this is not possible, then CMC discourse units should at least be anonymized. Moreover, an array of to some extent important tasks must be accomplished before the data can be integrated into a publicly accessible corpus. Typical „netspeak“ elements and phenomena of „written speech", such as the English *gotta* (< *got to*) and *dunno* (< *don't know*) or the German *haste* (< *hast du*), *ham* (< *haben*) and *willste* (< *willst du*; cf. 3.2.3), must be processed correctly by search tools. In order to deal with typing errors, one can possibly revert to spelling-tolerant search techniques as used e.g. in digital dictionaries. All in all, an array of non-trivial tasks must be accomplished before the data can be integrated into a publicly accessible corpus.

Preferably the construction and the processing of CMC corpora should henceforth be more strongly geared toward methods and standards from corpus linguistics. However, a current desideratum would be appropriate annotation standards that are suited to capture the specific discourse structures of interactive multiparty CMC genres, such as threads in discussion groups or chat-logfiles. Influential standards for corpus annotation, such as the guidelines of the *Text Encoding Initiative* (*TEI*), do not yet account for CMC genres. Such an extension would be very beneficial, especially for the development of specialized search tools. Up to now, many assumptions about the Internet's impact on language change have been based upon small datasets and a lot of intuition. Preferably, one should be able to compare CMC corpora, in which various CMC genres are contained in a balanced way, with corpora of other genres (e.g. newspaper articles, fiction, as well as oral discussions, interviews and counseling interviews). This would allow for empirically-based statements about language change due to digital media and the Internet and, furthermore, about how and with which effects people communicate in computer networks. It would be easier to investigate how the users linguistically adapt to the functions of CMC devices and how new patterns and genres emerge on the basis of existing patterns of verbal interaction. Our outline in Chapter 3 showed that some challenges still have to be overcome for this purpose. The growing relevance of the Internet for global communication shows that it is worthwhile to dedicate more attention to this area of corpus linguistics.

5. Literature

Androutsopoulos, Jannis K. (2003), Online-Gemeinschaften und Sprachvariation. Soziolinguistische Perspektiven auf Sprache im Internet. In: *Zeitschrift für germanistische Linguistik* 31(2), 173-197.

Beißwenger, Michael (2000), *Kommunikation in virtuellen Welten: Sprache, Text und Wirklichkeit.*

Stuttgart: ibidem.

Beißwenger, Michael (2003), Sprachhandlungskoordination im Chat. In: *Zeitschrift für germanistische Linguistik* 31(2), 198-231.

Beißwenger, Michael (2007), *Sprachhandlungskoordination in der internetbasierten Wissenskommunikation.* (forthcoming)

Beißwenger, Michael/Storrer, Angelika (eds) (2005), *Chat-Kommunikation in Beruf, Bildung und Medien: Konzepte - Werkzeuge - Anwendungsfelder*. Stuttgart: ibidem.

Berjaoui, Nasser (2001), Aspects of the Moroccan Arabic Orthography with Preliminary Insights from the Moroccan Computer-Mediated Communication. In: Beißwenger, Michael (ed), *Chat-Kommunikation. Sprache, Interaktion, Sozialität & Identität in synchroner computervermittelter Kommunikation. Perspektiven auf ein interdisziplinäres Forschungsfeld.* Stuttgart: ibidem, 431-465.

Bruckman, Amy/Erickson, Thomas/Fisher, Danyel et al. (2000), *Dealing with Community Data: A Report on the CSCW 2000 Workshop*. WWW resource: http://bulletin2.sigchi.org/archive/2001.4/comm_data.pdf. (4.9.2006).

Crystal, David (2001), *Language and the internet.* Cambridge University Press.

Danet, Brenda (1997), *Language, Paly and Performance in Computer-Mediated Communication. Final Report submitted to the Israel Science Foundation.* WWW resource: http://pluto.mscc.huji.ac.il/~msdanet/report95.htm. (4.9.2006).

Declerck, Thierry/Klein, Judith (1997), *Ein Email-Korpus zur Entwicklung und Evaluierung der Analysekomponente eines Terminvereinbarungssystems*. Paper, presented at the 6. Fachtagung der Sektion Computerlinguistik der Deutschen Gesellschaft für Sprachwissenschaft (DGfS/CL 97), Integrative Ansätze in der Computerlinguistik, 08.-10. Oktober, Heidelberg, Germany, 1997. WWW resource: http://www.coli.uni-sb.de/publikationen/softcopies/Declerck:1997:EKE.pdf (4.9.2006).

Döring, Nicola (2003), *Sozialpsychologie des Internet. Die Bedeutung des Internet für Kommunikationsprozesse, Identitäten, soziale Beziehungen und Gruppen.* 2. ed. Göttingen etc.: Hogrefe (Neue Medien in der Psychologie 2).

Evert, Stefan/Fitschen, Arne (2001), Textkorpora. In: Carstensen, Kai-Uwe/Ebert, Christian/Endriss, Cornelia, et al. (eds), *Computerlinguistik und Sprachtechnologie. Eine Einführung*. Heidelberg/Berlin: Spektrum, 369-376.

Feldweg, Helmut/Kibiger, Ralf/Thielen, Christine (1995), Zum Sprachgebrauch in deutschen Newsgruppen. In: *Osnabrücker Beiträge zur Sprachtheorie 50*: „Neue Medien", ed. by Ulrich Schmitz, 143-154.

Garcia, Angela Cora/Baker Jacobs, Jennifer (1999), The Eyes of the Beholder: Understanding the Turn-Taking System in Quasi-Synchronous Computer-Mediated Communication. In: *Research on Language and Social Interaction* 32(4), 337-367.

Grinter, Rebecca E./Palen, Leysia (2002), Instant Messaging in Teen Life. In: *Proceedings of the 2002 ACM conference on Computer supported cooperative work*, November 16-20, 2002, New Orleans/Lousiana.

Herring, Susan C. (1999), Interactional Coherence in CMC. In: *Journal of Computer-Mediated Communication* 4(4). WWW resource: http://jcmc.indiana.edu/vol4/issue4/herring.html. (4.9.2006).

Herring, Susan C. (2001), Computer-mediated discourse. In: Schiffrin, Deborah/Tannen,

Deborah/Hamilton, Heidi E. (eds), *The Handbook of Discourse Analysis*. Oxford: Blackwell, 612-634.

Herring, Susan C. (2004), Computer-mediated discourse analysis: An approach to researching online behavior. In: Barab, S. A./Kling, R./Gray, J. H. (eds), *Designing for Virtual Communities in the Service of Learning*. New York. Preprint: http://ella.slis.indiana.edu/~herring/cmda.html. (4.9.2006).

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

Herring, Susan C./Scheidt, Lois Ann/Bonus, Sabrina/Wright, Elijah (2004), Bridging the Gap: A Genre Analysis of Weblogs. In: *Proceedings of the 37th Hawaii International Conference on System Sciences*. WWW-Resource: http://csdl2.computer.org/comp/proceedings/hicss/2004/2056/04/205640101b.pdf . (4.9.2006).

Janich, Nina (2002), Von Lust und Leid. Metakommunikation in der E-Mail am Beispiel einer universitären Mittelbau-Initiative. In: Ziegler, Arne/Dürscheid, Christa (eds), *Kommunikationsform E-Mail*. Tübingen: Stauffenburg (Reihe Textsorten, Bd. 7), 217-243.

Jones, Rodney (2001), *Beyond the Screen. A Participatory Study of Computer Mediated Communication Among Hong Kong Youth*. Paper presented at the Annual Meeting of the American Anthropological Association Nov. 28 - Dec. 2, 2001. WWW resource: http://personal.cityu.edu.hk/~enrodney/Research/ICQPaper.doc (4.9.2006).

Journal of Computer-Mediated Communication (*JCMC*). Online-Journal. http://jcmc.indiana.edu/ (4.9.2006).

Journal of Interactive media in Education. Online-Journal. http://www-jime.open.ac.uk (4.9.2006).

Journal of Online Behaviour. Online-Journal. http://www.behavior.net/JOB/. (4.9.2006).

Language@Internet. Online-Journal. http://www.languageatinternet.de. (4.9.2006).

Leung, Louis (2001), College student motives for chatting on ICQ. In: *New Media & Society* 3(4), 483-500.

Leung, Louis (2002), Loneliness, Self-Disclosure, and ICQ („I Seek You") Use. In: *CyberPsychology & Behavior* 5(3), 241-251.

Maynor, Natalie (1994), The Language of Electronic Mail: Written Speech? In: Little, Greta D./Montgomery, Michael (eds), *Centennial Usage Studies*. Tuscaloosa 1994, 48-54.

Mehler, Alexander/Gleim, Rüdiger (2005): The Net for the Graphs – Towards Webgenre Representation for Corpus Linguistic Studies. In: Baroni, Marco/Bernardini, Silvia (Eds.): WaCky! Working Papers on the Web as Corpus. Bologna, 191-224.

Murray, Denise E. (1989), When the medium determines turns: turn-taking in computer conversation. In: Coleman, Hywel (ed), *Working with Language. A Multidisciplinary Consideration of Language Use in Work Contexts*. Berlin/New York: Mouton de Gruyter (Contributions to the Sociology of Languages 52), 319-337.

Naumann, Karin (2005), Kann man Chatten lernen? Regeln und Trainingsmaßnahmen zur erfolgreichen Chat-Kommunikation in Unterrichtsgesprächen. In: Beißwenger/Storrer (eds) (2005), 257-272.

Ogura, Kanayo/Nishimoto, Kazushi (2004), *Is a Face-to-Face Conversation Model Applicable to*

Chat Conversations? Paper presented at the Eighth Pacific Rim International Conference on Artificial Intelligence (PRICAI2004). WWW resource: http://ultimavi.arc.net.my/banana/Workshop/PRICAI2004/Final/ogura.pdf. (4.9.2006).

Ooi, Vincent B.Y. (2001), Aspects of computer-mediated communication for research in corpus linguistics. In: *Language and Computers* 36(1), 91-104.

Paccagnella, Luciano (1997), Getting the Seats of Your Pants Dirty: Strategies for Ethnographic Research on Virtual Communities. In: *Journal of Computer-Mediated Communication* 3(1). http://jcmc.indiana.edu/vol3/issue1/paccagnella.html. (4.9.2006).

Pankow, Christiane (2003), Zur Darstellung nonverbalen Verhaltens in deutschen und schwedischen IRC-Chats. Eine Korpusuntersuchung. In: *Linguistik online* 15. WWW resource: http://www.linguistik-online.de/15_03/pankow.pdf (4.9.2006).

Panyametheekul, Siriporn/Herring, Susan (2003), Gender and Turn Allocation in a Thai Chat Room. In: *Journal of Computer-Mediated Communication* 9(1). WWW resource: http://jcmc.indiana.edu/vol9/issue1/panya_herring.html. (4.9.2006).

Paolillo, John (1999), The Virtual Speech Community: Social Network and Language Variation on IRC. In: *Journal of Computer-Mediated Communication* 4(4). WWW resource: http://jcmc.indiana.edu/vol4/issue4/paolillo.html. (4.9.2006).

Porter, Constance Elise (2004), A Typology of Virtual Communities: A Multi-Disciplinary Foundation for Future Research. In: *Journal of Computer-Mediated Communication* 10(1). WWW resource: http://jcmc.indiana.edu/vol10/issue1/porter.html. (4.9.2006).

Rehm, Georg (2002), Schriftliche Mündlichkeit in der Sprache des World Wide Web. In: Ziegler, Arne/Dürscheid, Christa (eds), *Kommunikationsform E-Mail*. Tübingen: Stauffenburg (Reihe Textsorten 7), 263-308.

Rehm, Georg (2006): Hypertextsorten: Definition, Struktur, Klassifikation. Diss., Univ. Giessen. WWW resource (Giessener Elektronische Bibliothek): http://geb.uni-giessen.de/geb/volltexte/2006/2688/ (3.9.2006).

Reid, Elizabeth M. (1991), *Electropolis: Communication and Community on Internet Relay Chat*. WWW resource: http://www.irchelp.org/irchelp/misc/electropolis.html. (4.9.2006).

Runkehl, Jens/Schlobinski, Peter/Siever, Torsten (1998), *Sprache und Kommunikation im Internet. Überblick und Analysen.* Opladen/Wiesbaden: Westdeutscher Verlag.

Siebenhaar, Beat (2006), Code Choice and Code-Switching in Swiss-German Internet Relay Chat Rooms. In: Androutsopoulos, Jannis (ed.), Sociolinguistics and computer-mediated communication. Theme Issue, *Journal of Sociolinguistics 10*(4) (September 2006), 481-509.

Storrer, Angelika (2001), Getippte Gespräche oder dialogische Texte? Zur kommunikationstheoretischen Einordnung der Chat-Kommunikation. In: Lehr, Andrea/Kammerer, Matthias/Konerding, Klaus-Peter et al. (eds), *Sprache im Alltag. Beiträge zu neuen Perspektiven in der Linguistik. Herbert Ernst Wiegand zum 65. Geburtstag gewidmet*. Berlin: de Gruyter, 439-465.

Storrer, Angelika (2001a): Sprachliche Besonderheiten getippter Gespräche: Sprecherwechsel und sprachliches Zeigen in der Chat-Kommunikation. In: Beißwenger, Michael (ed): *Chat-Kommunikation. Sprache, Interaktion, Sozialität & Identität in synchroner computervermittelter Kommunikation. Perspektiven auf ein interdisziplinäres*

Forschungsfeld. Stuttgart: ibidem, 3-24.

Todla, S. (1999), *Patterns of communicative behaviour in Internet chatrooms*. Unpublished master's thesis, Chulalongkorn University.

van Eckert, Edgar (2005): Termingebundene Chats *one-to-one* in der psycho-sozialen Beratung. In: Beißwenger/Storrer (eds) (2005), 349-359.

Viegas, Fernanda B./Wattenberg, Martin/Kushal, Dave (2004), *Studying Cooperation and Conflict between Authors with history flow Visualizations*. Paper presented at the Conference on Human Factors in Computing Systems, Vienna. WWW resource: http://web.media.mit.edu/~fviegas/papers/history_flow.pdf. (4.9.2006).

Vilmi, Ruth (1994*), Global Communication through Email: An Ongoing Experiment at Helsinki University of Technology*. Paper presented at TESOL 94 Conference, Paris. http://www.hut.fi/~rvilmi/Publication/global.html. (20.6.2005).

Vronay, David/Smith, Marc/Drucker, Steven (1999), Alternative interfaces for chat. In: *Proceedings of the 12th annual ACM symposium on User interface software and technology* (CHI Letters 1,1), 19-26.

Werry, Christopher C. (1996), Linguistic and interactional features of Internet Relay Chat. In: Herring (ed.), 47-63.

Yates, Simeon J. (1996), Oral and written linguistic aspects of computer conferencing. In: Herring (ed), 29-46.

Zitzen, Michaela (2004), *Topic Shift Markers in asynchronous and synchronous Computer-mediated Communication (CMC)*. PhD-Thesis, Universität Düsseldorf. WWW resource: http://diss.ub.uni-duesseldorf.de/home/etexte/diss/file?dissid=771 (4.9.2006).

Zitzen, Michaela/Stein, Dieter (2004), Chat and conversation: a case of transmedial stability? In: *Linguistics 42*(5), 983-1021.

Computer-Mediated Communication (CMC) is the research field that explores the social, communicative and linguistic impact of communication technologies, which have continually evolved in connection with the use of computer networks. The main focus of CMC research is on Internet-based technologies and their genres: e-mail, mailinglists, discussion groups (forums and bulletin boards), Internet Relay Chat (IRC) and webchats, Instant Messaging (ICQ, AIM & Co.), MUDs, Voice-over-IP applications (Skype etc.), Web-based videoconferencing, weblogs and hypertext (incl. wikis). Types and examples of Computer-mediated communication (CMC) is defined as any communicative transaction that occurs through the use of two or more networked computers.[1] While the term has traditionally referred to those communications that occur via computer-mediated formats (e.g., instant messages, e-mails, chat rooms), it has also been applied to other forms of text-based interaction such as text messaging.[2] Research on CMC focuses largely on the social effects of different computer-supported communication technologies. Many recent studies involve Internet-based social networking supported by social software Request PDF | Building and Analysing Corpora of Computer-Mediated Communication | This chapter addresses problems encountered during the construction and analysis of a synchronic corpus of computer-mediated discourse. The corpus | Find, read and cite all the research you need on ResearchGate.Â Analysis focuses on data taken from a corpus of computer-mediated chat-room interaction. It investigates the ongoing performance of sexualised place (and place-based sexuality) through the use of language in online chat-rooms. The central questions focus on how the shared imaginary of a room helps to shape the performances of genders and sexualities and how the gendered and sexualised discourses sexualise the room. Computer-mediated communication CMC) is any form of communication between two or more individual people who interact and/or influence each other via separate computers through the Internet or a network connection - using social software. CMC does not include the methods by which two computers communicate, but rather how people communicate via computers. (Definition from the Wikipedia:Computer-mediated communication, feb 2006).