RiCL Research in Corpus Linguistics

New media, new challenges: exploring the frontiers of corpus linguistics in the linguistics curriculum

Nuria Hernández[1]
Universität Duisburg-Essen / Germany

Abstract – This paper introduces a new corpus of computer-mediated communication which is currently being compiled at the University of Duisburg-Essen. Based on the experience from this project, the paper also discusses the possibility of implementing major issues in corpus construction into the academic curriculum of young linguists in the form of project-based learning. A variety of new challenges and possible solutions regarding the compilation and processing of new media language are presented.

Keywords – blogs, CMC, DMC, emoticons, Facebook, image boards, new media, project-based learning, SMS, Twitter, YouTube

1. Introduction

In recent years, the study of language variation and change has extended to include a new and thrilling area of research: the language used in digital media. Research into 'digital discourse', 'computer-mediated communication', 'Internet language', 'Netspeak' and 'Textspeak' (Crystal 2004, 2006, 2010; see Herring 2007 for a more fine-grained classification scheme) is concerned with newer forms of correspondence, such as e-mails, chat, SMS or blogs, and more recently with the wide array of social media facilitating a fast and global exchange of user-generated content. While some people are concerned that the new technology may have a negative impact on language use, others argue the reverse, saying that the new media have encouraged a dramatic expansion in the creative capacity of language (for example Crystal 2006: 275).

The corpus presented in this paper provides an empirical basis upon which these assumptions can be tested. Despite the fact that linguistic research in computer-mediated communication (CMC) is growing at a fast pace, corpus-linguistic studies in the field often cite project-related corpora which are not readily accessible to the linguistic community (cf. Beißwenger and Storrer 2008). Examples would be the *CoSy corpus* (Yates 2001) or the *Swiss German Webchat Corpus* (Siebenhaar 2006). Databases for general use, on the other hand, include corpora as varied as, for instance, the *Dortmunder Chat-Korpus* (http://www.chatkorpus.uni-dortmund.de) or the *Enron Email Dataset* (http://www-

[1] I would like to thank all of my students who contributed to the first version of the DMC during the academic winter term 2011–2012 and thereafter. This project would not have been possible without their scrutinising questions, enthusiasm and their valuable input and ideas. For the 'Blogs' component: Linda Bleyer, Mark Elpers, Dominik Nebel, Yvonne Willuhn and Frauke Witt. For 'Image Boards': Sean Sams. For 'SMS German': Axel Bund, Seda Kiraç and Sebastian Krebs. For 'SMS English': Kate Roberts, Fiona Seward and Seán Upton. For 'Twitter': Catherina Hofmann, Melike Inan and Sabine Lange. For 'Facebook': Josua Ehmann, Lena Raue, Tina Terlinden and Nadine Vangenhassend. For 'Youtube': Julia Daitche, Shaun Hughes, Julian Kloss, Maria Laura Salerno and Ganna Strashnenko.

2.cs.cmu.edu/~enron). At present, however, no multi-genre corpus is available which could be used for research purposes as well as for the teaching of corpus-linguistic methods.

The need for a detailed linguistic investigation of current developments in CMC and the shortage of available data make a strong argument for a new corpus. This led to the project *Digital Media Corpus* (*DMC*), which was first presented at CILC2012, under the supervision of the current author. At the beginning of the project it was decided to realise this task within the framework of a graduate linguistics seminar in order to give students the chance to explore the world of corpus linguistics from a different angle, using a problem-oriented, project-based approach (Wrigley 1998; Stoller 2002). Viewed in this light, the lack of available data is an opportunity for exploring new frontiers.

The current paper reports on experience drawn from the *DMC* project which could prove useful for future experiments in the field, including the multiple challenges faced in the processing of different CMC genres, or 'socio-technical modes' (cf. Herring 2002; 'genre' and 'mode' will be used interchangeably in this paper). At present, the *DMC* comprises over 104,000 words from weblogs (blogs), image boards, SMS, *Twitter*, *Facebook* and *YouTube*, in English and German. The components differ in size and each component looks slightly different due to compositional differences between the source texts (e.g., 'text only' vs. 'text + pictures'; more details in section 3). Nevertheless, the preliminary version presented in this paper is a first milestone in working towards a consistently formatted database that will be made freely available for linguistic studies on CMC.

2. THE PROJECT

The *DMC* project began in the winter term of 2011, with an advanced seminar called 'Language in the New Media' for students in their third or fourth year of studies in English Literature and Linguistics (teacher education programme and bachelor's degree). The ultimate reason for including such a project in the linguistics part of the curriculum was to explore a different way of teaching corpus linguistics. An additional appeal of digital modes such as blogs, SMS, *Facebook* or *Twitter*, was that they are frequently used by the students and teachers themselves, often on a daily basis, thus adding a valuable emic perspective to their investigation. From a linguistic point of view, the spontaneity and the high level of emotivity in these modes promise a highly idiomatic and less self-monitored use of language which is difficult to elicit by other means.

While introductions to corpus linguistics tend to focus on the discussion and analysis of already existing databases, the aim of this seminar was to confront students more directly with the problems usually faced by corpus linguists themselves. Starting from scratch, the compilation of a completely new corpus provided ample opportunity for active discussion and decision-making. Unlike in previous seminars, the analysis of linguistic features was not realised in class, but was outsourced to subsequent term paper projects in order to allocate more time to the acquisition of corpus-compilation skills and data awareness. The results described in the following sections will therefore mainly refer to data compilation and processing; the advantages of the present approach for students analysing CMC language, and its comparability with other didactic approaches, will be touched on in section 5.

The overall time frame of the seminar consisted of weekly 90-minute classes in one of the university's computer pools, over a period of fourteen weeks. All participants had basic computer proficiency, but no previous experience in data collection or corpus compilation. After a general introduction to corpus linguistics, the theoretical issues relating to the changing nature of text (Ferrara, Brunner and Whittemore 1991; Crystal 2010), the different technical and social factors influencing CMC language (Yus 2011), as well as different CMC resources (websites, journals, dictionaries) in weeks 1 through 3, the project was divided into the following steps and goals, each under the guidance of the lecturer as the project supervisor.

- Planning and organisation (weeks 4 and 5)
  The first step consisted in the specification of the overall task and goals, along the lines of an "ill defined task with a well-defined outcome" (Capraro and Slough 2009), and the assignment of collaborative workgroups with up to 5 students per team, each group focusing on one CMC genre chosen by the students themselves; the only selection restrictions were copyright and privacy concerns (e.g., informed consent in the case of SMS); the result was a general corpus structure with 6 individual components; the individual teams decided how to proceed with collecting the respective data.
- Cooperation and creation (weeks 6 through 8)
  Raw data were collected by the different teams between the sessions, as an on-going home assignment, followed by partially supervised data processing in class (transcription and tagging); each seminar session started with an open discussion of issues relating to the textual markup and the tagging of special symbols and icons found in the different modes; step by step, a common tag list was generated for the entire corpus; a common text header format with text and user variables for all components was devised; the corpus was given a name.
- Control and reflection (weeks 9 and 10)

This stage consisted in mutual proofreading, feedback and correction of text files across the teams, in class and outside of class; the strategies chosen by each team were revised by another team, in some cases leading to major changes in the markup.

- End product to share (weeks 11 through 14)
  The project concluded with the collective writing of the corpus manual; the introductory part was written by the lecturer, and subchapters about the individual components were written by the student teams, once more in and outside of class; at the end of the seminar, students were offered the opportunity to further explore their data in a term paper focusing on the language used in the corpus, and to continue being involved in the corpus project.

In order to achieve the goals set out at the beginning, a variety of challenges had to be addressed, due in part to the fact that the corpus was being compiled from scratch in a self-motivating approach, and due in part to common issues in empirical linguistics, such as the protection of the authors' privacy.

The greatest challenge lay in formatting texts from different CMC genres. Although the up-and-coming research area of computer-mediated communication has attracted growing attention over the last few decades, linguistic publications and information on corpus design are still scarce (e.g., the electronic journal *Language@Internet*). For some modes, such as image boards, no corpora or linguistic studies existed at all, which put the respective students in the role of linguistic pioneers – in some cases enlivening their enthusiasm, in others deflating their confidence. Even well-researched modes, such as SMS or blogs, posed some open challenges. How should we tag special symbols and emoticons in a consistent, machine-readable format? What should we do with the many colloquial expressions, non-standard abbreviations and creative uses of language found in these new media? How, for example, would one tag a mixed-code expression such as *4tel 4 4* (German *viertel vor vier*; see section 4.7)? Should references to pictures and other websites be included in the transcripts? And, last but not least, how can user variables such as age, sex and origin be retrieved in media used by a largely anonymous global community? These are only some of the questions that had to be addressed. Before we look at possible solutions in more detail, a brief description of the corpus itself is in order.

3. THE *DMC* CORPUS

3.1. General structure

This section gives a brief introduction to the overall structure of the corpus and some special properties of its components. At present, the *DMC* contains approximately 104,200 words from 216 transcripts in 6 individual components: 'Blogs', 'Image boards', 'SMS' (English and German), '*Twitter*', '*Facebook* posts' and '*YouTube* comments'. For the time being, e-mails were not included because of the difficulties that this medium presents for the definition of 'text', due to partial text deletion, framing and intercalation of responses (cf. Crystal 2011). However, the multi-genre design of the corpus would allow a later inclusion, which could also make an intriguing topic for a future installment of the course.

The word counts of the individual components seen in Table 1 differ considerably, due to characteristic differences between genres (for example, short text messages vs. long text passages in 'Blogs').

COMPONENT	TEXT ID	TEXT FILES	WORD COUNT[2]
Blogs	BLG	3	30,800
Image boards	IMB	12	7,300
SMS, German	TXT…G	69	5,000
SMS, English	TXT…E	73	2,900
Twitter	TWT	7	43,800
Facebook posts	FBP	25	1,500
YouTube comments	YTC	27	12,000

Table 1. *DMC* components and word counts (June 2012)

Differences between the components also become visible in the directory structure. Figures 1 and 2, for instance, show the directory structures of the 'Blogs' and '*Twitter*' components. In 'Blogs', the folder for each blog contains a text file (tagged transcript), as well as the different pictures from the original website (also compare Figure 4 below), whereas '*Twitter*' contains text files only.

[2] Approximate word count, excluding text headers and tags.

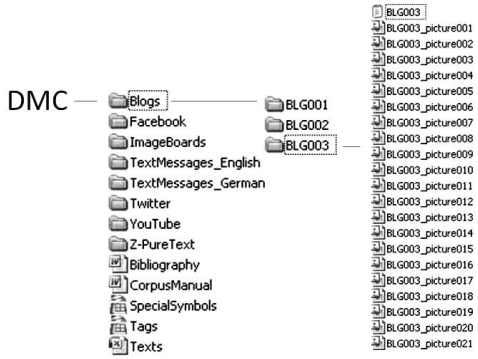


Figure 1. Directory structure, 'Blogs' component

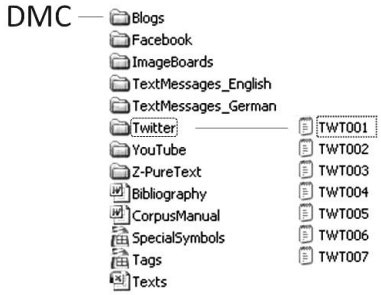


Figure 2. Directory structure, 'Twitter' component

During the collection process (November 2011 through January 2012), all data extracted for the different components were transformed into plain text files, and special symbols, icons and emoticons were marked with tags as seen in the examples below. Each transcript in the corpus was given a file name composed of a component ID (BLG for 'Blogs', IMB for 'Image boards', etc.), followed by a three-digit running number (BLG001, BLG002...), and each transcript was preceded with a text header containing the basic text and user variables.

3.2. The different components

3.2.1. Blogs

Weblogs, or blogs, are personal online journals of individual users or small groups which have enjoyed great popularity since the late 1990s. Unlike synchronous modes where all participants are online at the same time (e.g., Internet Relay Chat), blogs are less 'conversational' and, therefore, often perceived as closer to the written end of the written-spoken continuum (cf. Peterson 2011).

The current version of our corpus contains three different blogs with three different topics. Since the primary focus in the collection process was on language data, the topics were not a decisive factor; the students simply chose blogs they were familiar with.

Each blog transcript starts with a header containing the file name, the blog URL, the user's name, the language used, the posting time, the user's age and sex, and the general topic of the blog. Because of the many pictures occurring in the blog posts, and because of the fact that users frequently refer to the pictures in the text, it was decided that each blog be given its own folder containing the transcript as well as the corresponding picture files (compare Figures 1 and 4).

Blog	URL	Users	Age	Word count (approx. tokens)
BLG001	delicatehummingbird.blogspot.com	female	26	10,000
BLG002	gofugyourself.com	female, female	unknown, unknown	12,900
BLG003	dooce.com	female	36	7,900

Table 2. Blogs in the *DMC* (June 2012)

3.2.2. Facebook *posts*

Launched in 2004, *Facebook* has become the most popular social network worldwide. According to information provided on *Facebook*'s website, over 650 million people are said to be currently using the network on a daily basis (*Facebook* 2013a). Its mission is "to give people the power to share and make the world more open and connected" (*Facebook* 2013b). *Facebook* users may upload pictures, share links and videos and connect with friends all over the world. All users can comment on any content added by their friends, a special feature being the option to signal approval of another user's comment or content by giving it a 'thumbs up'.

The data collected for the *DMC* consists of comments which the students themselves, as *Facebook* users, had previously posted in reply to other users' status reports. Each transcript presents one 'conversation,' starting with a status update by one user and the subsequent posts responding to this update (see example (3)). Threads which contained links or pictures were not included. Since status reports basically describe what is on the user's mind, some posts can be confusing or do not seem to make much sense to someone who is not immediately involved in the exchange. The reader of a *Facebook* post does not necessarily know the context of the respective entry and commenters are in no way obliged to explain themselves.

The current version of the *DMC* contains 24 *Facebook* transcripts in German, but other languages, including English, could be added at any time.

3.2.3. Image boards

Image boards are a kind of bulletin board system, much like a public chat room, where users can create threads on different topics. Originally invented in Japan, image boards have been copied in other countries, especially in the United States. The most famous image board at present is *4chan*, which stars among the top 900 most visited websites with up to 450,000 postings per day. The main language in image boards is English, but any user may start a thread in another language.

The hallmark of this medium is its total anonymity. All image board users are anonymous, to the extent that even nicknames are avoided, and anybody can read any uploaded post. Instead of official registration, image boards use tripcodes which contain no user details. In addition, the threads are extremely short-lived and often deleted after one or two hours, making them the least persistent contributions with the, assumedly, least meta-linguistic awareness in the corpus (cf. Herring 2007: 15). By saving the data, our project breaches this policy to some extent, but anonymity remains guaranteed in the transcripts.[3]

Currently, the 'Image boards' component of the *DMC* contains 12 text files with over 7,300 words. In this mode, too, posts are often accompanied by pictures which comment on the written text in some way. In fact, discussions are highly graphic-centric, often initiated by posted images which can have follow-up pictures posted as responses. Researchers should note that these threads are possibly incomplete, since posts can be deleted after the image limit has been reached and extremely long threads were only partially extracted.

3.2.4. SMS

For SMS, as for most of the other modes described in this paper, no linguistic corpus was publicly available when the project started. So far, this component contains messages in English and German, with the addition of further languages being planned. The total word count currently amounts to almost 5,000 for German, and 2,900 for English (excluding text headers and tags). A first example of the brief messages sent between (mobile) phones and other devices is shown in Figure 3, followed by further examples below.

SMS are usually short, and individual exchanges do not go on for very long. Together with *Facebook* posts, these data are the most difficult to obtain, since they are generally perceived as more personal than other CMC modes.

---

[3] Complaints against the use of these data should be directed to the author; they will be taken seriously.

<Simone> Hey Anja, ja war leider in der uni.. Bin total fertig! Ich hoffe zumindest, dass ihr gestern alle spass hattet! [reg=xxx] kisses [\reg]

Figure 2. Original SMS on mobile phone screen, plus transcript (*DMC*, TXT009G)

3.2.5. Twitter

The social networking service *Twitter* was created in 2006 as a medium for keeping in touch with both friends and the general public. *Twitter* enables its users to send and read text-based posts of up to 140 characters, known as *tweets*. The character limit was imposed to interface easily with text messaging services. In the last few years, this medium has been increasingly used by celebrities who enjoy regular contact with their fans and supporters, including singers, actors and politicians. This is also reflected in the *DMC* '*Twitter*' component, which contains original data from seven different *Twitter* accounts of various singers, such as Katy Perry's and Bruno Mars's. Each file in this component contains the tweets of the account owner and comments by various commentators (answers to the original tweets).

In order to use *Twitter*, one has to set up an account, including a username (usually a nickname) and a profile picture. *Twitter* is hence slightly less anonymous than the above-mentioned image boards and even age and gender are occasionally provided in the commentator profiles.

Collecting data for this medium was relatively easy—a fact which is reflected in the highest word count of 43,800; see Table 1.

3.2.6. YouTube *comments*

YouTube, a video-sharing website created in 2005, is the first address for many Internet users looking for free videos and music, including those who also want to share their thoughts and impressions with a larger community. *YouTube* language has been severely criticised as "[j]uvenile, aggressive, misspelled, sexist, homophobic" (Owen and Wright 2009), but so far such assumptions have not been tested on any empirical grounds. Corpora like the *DMC* can help close this gap.

In the corpus, the audiovisual material itself is not included, the focus being on the concurrent user comments. Assuming that comments on different topics might differ linguistically, the student team decided to include a range of topics in order to give a more balanced picture of *YouTube* language. At present, the '*YouTube*' component contains 27 different files with 6 different topics selected from the large variety discussed online: music, education, comedy, babies, politics and news stories. A first example of a '*YouTube*' file is shown in (4).

4. CHALLENGES AND RESULTS

4.1. User privacy

The first challenge that the students were confronted with during data collection concerned the users' privacy. The protection of user (speaker/author) privacy is a well-known issue in empirical linguistics, concerning especially those genres where the users themselves decide how much private details they give out and with whom they want to share their thoughts.

Two components in our corpus are especially affected by this issue: 'SMS' and '*Facebook* posts'. In these modes, most of the data was contributed by the team members themselves, i.e., their own text messages and posts from their own *Facebook* accounts, in agreement with the respective co-users. Despite the fact that the project was conducted in the Department of Anglophone Studies, this procedure resulted in both an English and a German SMS subcomponent, and predominantly German *Facebook* posts (which will hopefully be extended to English in the future).

As an additional protective measure in both components, the names of users who were not part of the research teams were made anonymous, and some messages or fragments of text which were considered to contain very personal information were deleted. Other user variables were kept, as seen in Table 3.

The privacy issue does not only concern the usernames. In any online genre there are users who prefer not to disclose their personal details, which makes the user variables less reliable than in other types of linguistic data. Especially the 'age' variable should always be taken with a grain of salt. It is virtually impossible to know how much one can trust the information extracted from the Internet, 'age' being particularly unreliable. In extract (4), for example, the *YouTube* user BeraSk8, one of the commentators on US rapper Dr Dre in YTC020, purports to be 111 years old—and he is only one of many alleged 100+ users on *YouTube*.

Before we continue with the next challenge, here are some examples of transcripts from different parts of the corpus. In German examples, the English translations are given in italics.

(1) SMS transcript, German
<text ID TXT016G> <language German> <date 112011>
<user Philip,Marvin> <user age 26,25> <user sex m,m> <native language German,German>

<Philip> Hi Philip. Kann ich morgen deinen Ghettoblaster ausleihen?
<Philip> Hi Philip. Can I borrow your ghetto blaster tomorrow?

<Marvin> das tut mir leid der ist ja nicht von mir sondern von unserem Team [reg = u] und [\reg] zurzeit nicht in meiner Gewalt!
<Marvin> I'm sorry it doesn't belong to me but to our team and it's currently not under my thumb!

<Philip> [reg = aso] ach so [\reg]. Dachte wäre deiner. OK
<Philip> I see. Thought it was yours. OK

(2) SMS transcript, English
<text ID TXT003E> <language English> <date 102011>
<user Sean,Barry> <user age 20,21> <user sex m,m> <native language English,English>

<Sean> some burn on the rugby but on the other hand we're all off to poland

<Barry> some burn [reg=alrite] alright [\reg] haha. what you going there for? train ya? what part you going to?

<Sean> man for the euros in the summer!!

<Barry> haha ya man it's all about the soccer team. they'll probably get [reg=bate] beaten [\reg] by armenia the way things are going.

(3) *Facebook* posts, German
<text FBP007> <language German> <posting time 112011>
<user Kordula,Becky,Zack,Mira,Gordon,Carla> <user sex f,f,m,f,m,f> <user age ,,,,,>

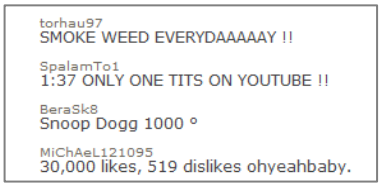
<Kordula> [26/11/2011 1:35pm]
ich will ans [emphcap] MEER [/emphcap]!!!!! Dicke Jacke, Gummistiefel, Schal, Mütze, Taschentücher, Geld für nen heißen Kakao und ab [reg=geeeehts] geht's [\reg]!
I want to go to the SEA!!!! Thick jacket, wellingtons, scarf, cap, tissues, money for a hot chocolate and off we go!

<Becky> [26/11/2011 1:36pm]
boah [reg= joo] ja [\reg], [reg= dat] das [\reg] [reg=wärs] wär's [\reg]
wow yeah, that would be great

<Zack> [26/11/2011 1:42pm]
wann [reg= solls] soll's [\reg] los gehen?
when do you want to go?

<Kordula> [26/11/2011 1:42pm]
hmmm. in [reg=ner] einer [\reg] stunde [em laugh] :D [\em laugh]
hmmm. in an hour :D
…..

(4) *YouTube* transcript plus screenshot



<text ID YTC020>
<topic music Dr Dre>
<language English>
<posting time 122010>
<user SpalamTo1,BeraSk8,…>
<user age 21,111,…>
<user country Poland,Brasil,…>
<URL http://www.youtube.com/watch?v=ejUARfOR7hE&feature=relmfu>

<user SpalamTo1>
<posting time 141210>
1:37 [emphcap]ONLY ONE TITS ON YOUTUBE[\emphcap]!!

<user BeraSk8>
<posting time 141210>
Snoop Dogg 1000[sym=°] degrees [\sym]
…..

4.2. Defining textual units

In his study "O brave new world, that has such corpora in it!", David Crystal remarks that, "[i]f there's one thing that unites all of us, in the field of corpus linguistics, it is that we assume we know a text when we see one" (Crystal 2011: 1). Unfortunately, but also intriguingly, this assumption is difficult to maintain when dealing with CMC genres like the ones discussed here. The traditionally definable properties of 'text'—such as spatial and temporal boundaries, and permanence—are hard to apply to newer media. The "stable, familiar, comfortable world" that corpus linguists once dealt with has changed, and research in digital discourse needs to rethink the notion of 'text' (Crystal 2011: 1).

In more concrete terms, we have to decide what to do with extra-textual elements, such as pictures, and textual elements that are not part of the main text or lead us to other texts, such as hyperlinks. During the compilation process, it soon became clear that the answers to these questions might vary, but the main criterion agreed upon by all student teams was that elements (both textual and extra-textual) should be included if, and only if, they are referred to in the main body of the text. In that respect, hyperlinks form part of the running text, but the texts to which they link do not.

Another question that had to be answered was at what point to cut off texts which lack the above-mentioned boundaries. Genres such as *Twitter*, for instance, have threads that can go on for a long time, often with extended intervals and, in most cases, these threads will continue after the collection of data for our project has ended. For the '*Twitter*' component, entire threads were obtained by clicking the 'all comments' view, on one specific date which is mentioned in the text header. This way, any thread could be chronologically extended in follow-up versions of the *DMC*.

4.3. Texts and pictures

In CMC genres such as blogs and image boards, pictures are regularly used to illustrate and comment (often humorously) or simply add visual impressions to the written text. In the texts themselves, these pictures are not always mentioned, but the connection is usually apparent. It was therefore decided, in both components, to include the pictures in the respective folders (see Figure 1), and to mark the original position of each picture with a *picture tag*.

The example in Figure 4 was taken from an American blog by an English native speaker, called dooce.com. The topics of this blog revolve around the author's everyday life, experiences and thoughts. Dooce.com has received numerous Weblog Awards for 'Best American weblog' (2005, 2008), 'Best-designed weblog' (2008), 'Weblog of the year' (2008), 'Most humorous weblog' (2005), 'Best writing of a weblog' (2005), and 'Lifetime achievement' (2008). In this example, the author writes about her dog, including pictures of him on the website. In the corpus transcript, these are indexed by consecutively numbered picture tags.

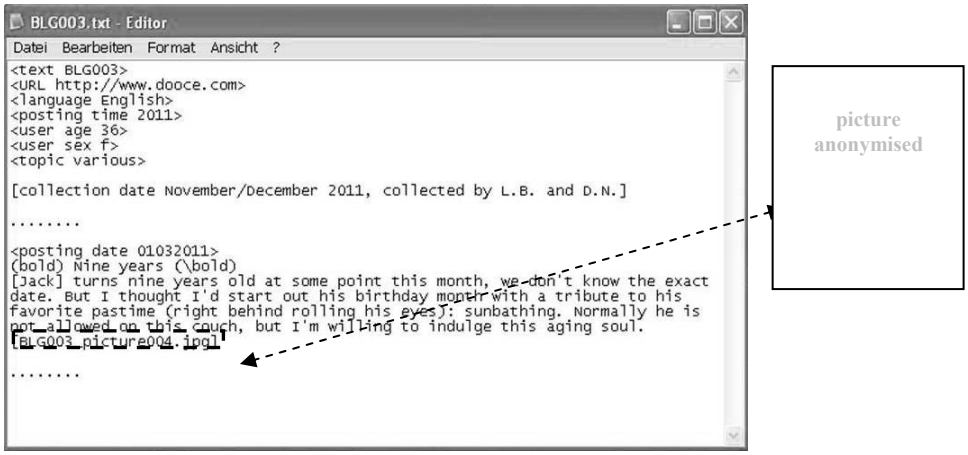


Figure 3. Blog transcript with picture tag (BLG003_picture004.jpg)

4.4. Consistent formatting

One of the goals of this project was to compile a consistently formatted CMC corpus for comprehensive analyses of new media language. In order to achieve this goal, multiple decisions had to be taken, once the data had been collected, in order to transfer them into a homogeneous format – always taking into account that the students had little or no experience in data processing.

First of all, it was decided that each transcript should be preceded by a header containing the basic text and user variables (using empty spaces for missing values). Due to differences in the accessibility of these variables, their number varies between the different genres, as shown in Table 3. In genres such as *Twitter* or *YouTube*, the personal details of the users are mostly unknown and cannot be deduced from the usernames (nicknames). The most anonymous genre –'Image boards' – naturally has the fewest variables; and '*Facebook* posts' is not specified for 'topic'. In *Twitter*, an additional distinction was introduced between the main author, i.e., the account holder (*user*), and the authors of other tweets, who are referred to as 'commentators'.

After determining the format of the headers, the issue of texts was addressed. Unlike more traditional genres, the ones included in this corpus exhibit features which compensate for prosody and other paralinguistic features typically associated with speech (see Crystal 2003: 291–293). In the texts this is, for instance, indicated by the use of emoticons (compensating for facial expressions and gestures), non-standard spellings (dialect features, slang, abbreviations), and different typographical conventions used to signal emphasis or a raised tone of voice. Other phenomena, such as the use of politically incorrect language and the frequent occurrence of orthographic mistakes, are linked to the spontaneity and the reduced level of formality in CMC. The different linguistic features tagged in the texts are described in the following sections.

BLOGS	IMAGE BOARDS	SMS	*TWITTER*	*FACEBOOK*	*YOUTUBE*
text ID	text ID	text ID	text ID	text ID	text ID
user		user	user	user	user
language	language	language	language	language	language
posting time	posting time	date	posting time	posting time	posting time
user age		user age	user age	user age	user age
user sex		user sex		user sex	
		native language			user country
topic	topic		topic		topic
			commentators		
URL	URL		URL		URL

Table 3. Text and user variables in the different *DMC* components

4.5. How to tag special symbols, icons and emoticons

The spontaneity, the relatively low level of formality, and the celerity with which language is used in digital discourse explain the frequent use of special characters and icons instead of fully written words. In the corpus files, we find

special symbols replacing words such as ‘and’ (&) or ‘degrees’ (°, see example (4)), different currency signs such as ‘dollar’ ($), the heart symbol used for ‘love’ (♥, also represented as “<3”), the *at* sign @ used for ‘directed at’ as well as local or temporal *at* in tweets and posts (English and other languages, too), and many more. The *Facebook* example in (10), for instance, shows how @ can be used to address more than one person in one post.

In all text files, special symbols were tagged with *symbol tags* in the way shown in examples (5)–(10) (once more, English translations are given in italics). This way, the symbols themselves are preserved in the text files whenever possible, but their transliteration into words is also given in order to facilitate word searches with text-based concordancers.

(5) <jasminlopez1970>
i will pray for you [reg=guyz] guys [\reg]..... just believe [sym=&] and [\sym] pray....
(*DMC*, YTC017)

(6) <katyperry 02112011>
It's [emphcap] HERE [\emphcap]! Grab [reg=urself] yourself [\reg] a horchata [sym=&] and [\reg] a churro cause the [sym=#] hash [\sym] CALIFORNIADREAMSTOUR goes to Mexico!
(*DMC*, TWT002)

(7) <Kordula> ... liebend gern, liebe Carla [sym=<3] heart [\sym]
<Kordula> ... I’d love to, dear Carla [sym=<3] heart [\sym]
(*DMC*, FBP007)

(8) <Timrath> Are they obliged to touch the dispatch box when they're speaking?
<Neighb0urhoodWatch> [sym=@] at [\sym] Timrath No
(*DMC*, YTC007)

(9) <MorseCoach> [em lol] lol [\em lol] guy sleeping [sym=@] at [\sym] 2:00
(*DMC*, YTC007)

(10) <Johannes> [12/12/2011 06:17pm]
[sym=@] at [\sym] flo: denkst wie dein [fl language] [reg=bro] brother [\reg] [\fl language] nur ans saufen [em crack up] xD [\em crack up]
[sym=@] at [\sym] flo: you always think of nothing but booze like your [reg=bro] brother [\reg] [em crack up] xD [\em crack up]
[sym=@] at [\sym] kathrin: ich bin grade in wien und mach ein praktikum für mein studium
[sym=@] at [\sym] kathrin: I’m in vienna just now doing an internship for my studies
(*DMC*, FBP012)

A regular strategy that users apply in order to reinforce the written comments and express their mood and emotions is the representation of facial expressions by so-called *emoticons*. Digital discourse is notorious for the use of predefined sequences of punctuation marks, such as the *smiley* :), which many programmes recognise and automatically convert into the corresponding pictorial representations (☺).

In the current version of the corpus, there are 37 different emoticons which had to be tagged accordingly, and more types will certainly appear as the corpus grows. Since each emoticon signals a different mood or facial expression, it was decided that each meaning would have to be specified within the *emoticon tag*.

Examples (11)–(14) show just a few occurrences of the many emoticons found in our data. Figure 5 shows the beginning of the extensive tag list that accompanies the corpus files.

(11) <calisunluvr> Awe. So cute and funny. [em smile] :) [\em smile]
(*DMC*, YTC023)

(12) <esa2go> the boy at the end could be little sheldon cooper [em laugh] :D [\em laugh]
(*DMC*, YTC027)

(13) <BrunoMars 15112011>
Damn we hit 5 Million! […] [reg=Lets] Let’s [\reg] take our shirts off [em nyah] :p [\em nyah]
(*DMC*, TWT003)

(14) <QuercusSola>
[sym=@] at [\sym] simbaglare714 I know what you mean. When I talk to neighborhood kids I have to switch to “dumb english”… [em lol] lol [\em lol] [em laugh] :D [\em laugh]
(*DMC*, YTC001)

1	Text/Symbol	Tag
2	lol	[em lol) lol [\em lol]
3	:)	[em smile] :) [\em smile]
4	:(	[em sad] :([\em sad]
5	;)	[em wink] ;) [\em wink]
6	:-)	[em smile nose] :-) [\em smile nose]
7	:-(	[em sad nose] :-([\em sad nose]
8	;-)	[em wink nose] ;-) [\em wink nose]
9	(-:	[em smile left] (-: [\em smile left]
10	|-)	[em hee] |-) [\em heehee]
11	|-D	[em hoo] |-D [\em hoho]
12	:->	[em smirk] :-> [\em smirk]
13	:-(	[em boohoo] :-([\em boohoo]
14	:-<	[em realsad] :-< [\em realsad]
15	:-|	[em hmm] :-| [\em hmm]
16	:-O	[em uhoh] :-O [\em uhoh]
17	:-o	[em shock] :-o [\em shock]
18	:-p	[em nyah nose] :-p [\em nyah nose]
19	:p	[em nyah] :p [\em nyah]
20	|-P	[em yuck] |-P [\em yuck]
21	:-X	[em lipssealed] :-X [\em lipssealed]
22	<:-)	[em dunce] <:-) [\em dunce]
23	>:-<	[em mad] >:-< [\em mad]
24	:-@	[em scream] :-@ [\em scream]
25	:-\	[em undecided] :-\ [\em undecided]
26	:-U	[em sarcastic] :-U [\em sarcastic]
27	:-D	[em laugh nose] :-D [\em laugh nose]
28	^ ^	[em grin] ^ ^ [\em grin]
29	B-)	[em cool] B-) [\em cool]
30	:D	[em laugh] :D [\em laugh]
31	XD (or xD)	[em crack up] XD [\em crack up]
32	;P	[em cheeky wink] ;P [\em cheeky wink]
33	-.-	[em irritated) -.- [\em irritated]
34	;D	[em wink laugh] ;D [\em wink laugh]
35	>.<	[em frustrated] >.< [\em frustrated]
36	=D	[em happy laugh] =D [\em happy laugh]
37	(=	[em happy smile left] (= [\em happy smile left]
38	O.O	[em shock] O.O [\em shock]
39	👍	[thumbs up]
40	👎	[thumbs down]

Figure 4. Screenshot of the first part of the *DMC* tag list

4.6. How to tag non-standard spellings, errors and abbreviations

In order to mark the great number of non-standard spellings and abbreviations found in the data, a special *regularisation tag* was introduced which permits to both keep the original item and insert a standardised variant. This way, dialectal, informal or slang realisations can be searched for directly or via the corresponding standard lexeme.

The examples shown in this section include all kinds of well-known phenomena which are usually associated with speech, such as the dropping of final *–g* in (15), but also creative innovations, such as the use of numbers or individual letters for homophonous words in (18), or extensive abbreviations which are not necessarily known outside a specific user community, as seen in (19). The latter are an especially frequent feature of image boards, which contain multiple abbreviations and figures of speech not found in other CMC genres. Two common abbreviations in image boards – *mfw* ("my face when") and *op* ("original poster") – are shown in (20) and (21), respectively.

The regularisation tag turned out to be one of the most frequent tags in the entire corpus; these are just a few examples.

(15) <beckymefford> So [reg=Freaken] Freaking [\reg] cute!!!!!! (*DMC*, YTC023)

(16) <justkidding93>And I'm a [reg=preachers] preacher's [\reg] kid to boot!
(*DMC*, YTC001)

(17) <Sean> some burn on the rugby but on the other hand we're all off to poland
<Barry> some burn [reg=alrite] alright [\reg] haha.
(*DMC*, TXT003E)

(18) <jasonderulo>
I'm excited [reg=2] to [\reg] [reg=b] be [\reg] going home [reg=4] for [\reg] thanksgiving! [reg=4] four [\reg] [reg=yrs] years [\reg] since I've enjoyed home [sym=&] and [\sym] [reg=fam] family[\reg] on t-day, not [reg=2] to [\reg] mention last [reg=yr] year [\reg] [sym=@] at [\sym] Ruby tuesday's! ha!
(TWT001)

(19) [reg=tihilw] this is how it looks worn [\reg] [BLG001_picture158.jpg]
(*DMC*, BLG001, referring to a picture in the blog)

(20) <No.2268571> [16:40] [pic 1324417255.jpg]
[reg=mfw] my face when [\reg] i see the body artist tucked in there
(*DMC*, IMB008)

(21) <No.2268577> [16:43] [2268564]
Have you seen the movie fight club [reg=op] original poster [\reg]?
(*DMC*, IMB00)

4.7. How to tag foreign language expressions

Another common feature in digital discourse are switches between languages. In our corpus, we found English words in German texts, Spanish words in English texts, and various other combinations. It was decided to mark these words in order to facilitate, for instance, the analysis of code-switching. The tag used here is a *foreign language tag* opening with the bracket [fl *value*], where *value* is the respective language of the tagged word or words. Foreign language expressions in our data range from individual words, as seen in the two German SMS in (22) and (23), to short phrases, such as (24), or even entire sentences, as seen in (25).

(22) <Christian> Wann seid ihr da, [fl English] guys [\fl English]?
When will you be there, guys?
(*DMC*, TXT105G)

(23) <Nena> Jetzt [reg=hab] habe [\reg] ich schon fast alle apps gelöscht [reg=nen] einen [\reg] viren scan gemacht und die scheiße schickt immernoch [fl English] fake [\fl English] nachrichten raus.
I have already deleted most of my apps, did a virus scan and this shit is still sending fake messages.
(*DMC*, FBP025)

(24) Tonight I am a Glamour magazine World's Most Beautiful All-Star Something-Something, and lovers, nobody deserves it [fl Spanish] mas que yo [\fl Spanish].
... more than me.
(*DMC*, BLG002)

(25) <anna10797> I defy anyone to say this lady [reg=isnt] isn't [\reg] talented! Feckin awesome!!
<diegohugostoso1> please! [fl Portuguese] alguém sabe o nome da primeira musica que ela cantou ??? [\fl Portuguese] thanks
... Does anybody know the name of the first song she sang?...
(*DMC*, YTC008)

Alongside such simple examples, we frequently find foreign language expressions which exhibit additional features requiring other tags. Just like any other passages in the discourse, interjections in a different language can contain non-standard spellings and abbreviations, and they can use the same typographical conventions, for example in order to signal emphasis as described in section 4.8. A combination of features can simply be marked by *nested tags*, as shown in (26) and (27) (repeated from (10)).

(26) Well Played, Jennifer Lopez "[fl Spanish] [emphcap] HOLA [\emphcap] [\fl Spanish]… [italics] sniffle [\italics]… [emphcap] LOVERS [\emphcap]".
(*DMC*, BLG002)

(27) <Johannes> [12/12/2011 06:17pm]
[sym=@] at [\sym] flo: denkst wie dein [fl language] [reg=bro] brother [\reg] [\fl language] nur ans saufen
[sym=@] at [\sym] flo: you always think of nothing but booze like your [reg=bro] brother [\reg]
(*DMC*, FBP012)

One of the most creative examples in the *DMC* is the mixed-code expression shown in (28), where German *viertel vor vier* 'quarter to four' becomes *4tel 4 4*. The author, *Philip*, uses digits instead of numbers to type in the time when he wants to meet. *Vier* 'four' becomes *4*. In addition, the preposition *vor* 'before/to' is represented by an English *4*, which is possible because of the near-homophony of the two expressions *vor* /fɔːɐ/ - *four* /fɔː(r)/. Note that the German word for number 4 is *vier* /fiːɐ/. While the first, second and fourth *4* in this message are pronounced in German, the third *4* must get the English pronunciation in order to make sense.

(28) <Philip>
Sorry aber das wird [reg=nix] nichts [\reg]. Sina kann erst doch um [reg=4] vier [\reg]. Also kannst länger arbeiten [em smile] :) [\em smile] bin [reg = 4tel] viertel [\reg] [reg=4] vor [\reg] [reg=4] vier [\reg] bei dir.
Sorry but I can't. Turns out Sina can only make it at 4. So you can work longer [em smile] :) [\em smile] I will be at your place at quarter to 4.
(*DMC*, TXT020G)

4.8. Typographical conventions signalling emphasis

Among the unique features that distinguish speech from writing is the use of prosodic elements, including emphasis through tempo and loudness (cf. Crystal 2003: 291). Different CMC genres have found a way to replace these elements by means of typographical conventions indicating increased emotivity and intensity. Overall, the two most widespread strategies – in texts which do not allow any other type of formatting – involve the use of capitalisation ([emphcap]) and asterisks ([emphast]), as shown in (29)–(33). Another convention which is found less frequently, additional spacing between letters ([emphspa]), has not occurred in our dataset so far.

(29) Something you collect: Monster bottle caps. Knowledge [emphcap] YEAH [\emphcap].
(*DMC*, IMG002)

(30) <katyperry 01112011>
Truly! RT [sym=@] at [\sym] Oh Ferras: tonight i'm going to wear…. [emphcap] NO MAKE UP! [\emphcap] best way to give [reg=y'all] you all [\reg] a fright!
(*DMC*, TWT002, RT 'retweet')

(31) <beautifulgirl95100>
[em lol] LOL [\em lol].[emphcap] EVERYBODY, THE FINEEE GUY AT THE END ON THE MOTORCYCLE, IS MICHAEL JACKSON'S NEPHEW, SIGGY JACKSON. [\emphcap]
< ShizzleKizzle07>
I will [emphcap] NOT [\emphcap] cry. [emphast] *sniffles and clears throat* [\emphast]
(*DMC*, YTC011)

(32) <mhairicatherine>
i [emphcap] LOVE [\emphcap] this!!!! seen it so many times and its utterly adorable
(*DMC*, YTC022)

(33) <SuperPeaceout14>
i love this video [em lol] lol [\em lol]
[…]
<zomgseriosuly>
Anthony Padilla? [emphast] *hungry face* [\emphast]
(*DMC*, YTC005)

Depending on the user interface, words can also be emphasised through a modification of the font, i.e., they can be underlined or set in bold type or italics—typographical conventions which are well known from writing. Since these changes are not displayed in plain text, we decided to mark them with the corresponding tags seen in Table 4.

Note that in the preliminary version of the *DMC* the use of emphatic asterisks is only found in *YouTube* posts, but it would probably not be restricted to this medium in a larger dataset.

Graphological conventions	*DMC* tags
capital letters used for emphasis/ shouting	[emphcap] … [\emphcap]
asterisks for emphasis	[emphast] … [\emphast]
letter spacing used for emphasis/ "loud and clear"	[emphspa] … [\emphspa]
underlined words	[underlined] … [\underlined]
words in italics	[italics] … [\italics]
words in bold type	[bold] … [bold]

Table 4. Graphological conventions signalling emphasis

4.9. Politically incorrect language: to tag or not to tag?

The final challenge in this project was the frequent use of swearwords and expletives, for instance in media such as *YouTube*, *Twitter* and image boards. It soon became apparent that most of the students involved felt uncomfortable including these words in the corpus without comment. Several solutions were proposed for tagging words such as *fucking*, *damn* and the like, but in the end it was agreed that, from a matter-of-fact linguistic perspective, there is no reason why these words should be distinguished from non-expletives.

In the future, the frequency of expletives in CMC, as compared to more traditional media, will certainly arouse some interest, and considering the widespread prejudices against certain CMC genres, this topic is in dire need of linguistic investigation, both qualitative and quantitative. It might therefore, at some point, make sense to introduce *expletive tags* in datasets such as the *DMC*.

5. Conclusion and outlook

The project presented in this paper proved to be a very positive experience, both from a didactic and from a corpus-linguistics point of view. Multiple challenges that were brought up during the collection and processing of the data were readily accepted by the students involved. Despite their lack of corpus and tagging experience, the students' familiarity with the genres at hand and the awareness that they could actively contribute to the production of "something new", more than outweighed the technical difficulties which are to be expected in this type of linguistic spadework.

On the basis of the continuous assessment of the tasks described in section 2 and the final course evaluation, it can be concluded that the students developed a firm understanding of human communication and of the differences between the various media and genres used to transmit information. In addition, the practical tasks in this seminar required particularly strong interpersonal skills. The communication and collaboration within the research teams provided an incentive for developing solutions in joint effort, completing assigned tasks within a given time frame, taking common decisions and sharing experiences.

As a final common task, the entire class wrote a corpus manual, comprising a general description of the textual markup and processing guidelines (written by the lecturer), as well as individual sections explaining the different components and their special characteristics. These sections were written by the students themselves – a task which proved more demanding than expected. Compared to the usual essays and term papers that students have to write during their studies, corpus manuals present a different genre with a very technical style and purpose. Thus, writing the manual presented an additional challenge and learning experience.

A strong motivation in this particular seminar was to create an "end product to share" (as mentioned in section 2), i.e., the corpus itself, which the course participants could subsequently use as an empirical basis for their own investigations. Regarding the student papers that resulted from the seminar, it is admittedly difficult to assess to what extent the didactic approach adopted in this project may have factored into the quality of the linguistic analyses. However, the general feedback from students who decided to write a term paper suggests that they felt more comfortable analysing data which they knew, and their newly obtained certitude as researchers who had been involved in the decision making and construction of their own database was positively reflected in how they construed their arguments in favour of the methodology and approach they chose for their investigations. A most encouraging response from various participants was their interest to continue contributing to the corpus afterwards.

In order to objectively assess the didactic value of the approach described in this paper, and in order to estimate its influence on student efficiency in corpus use, a special experiment would need to be designed to warrant the comparability with other corpus linguistic seminars. This could be implemented through a series of parallel or consecutive seminars on corpus linguistics using different didactic approaches for student groups with comparable computer skills and corpus experience.

With respect to the challenges discussed in section 4, the solutions proposed by the student teams were surprisingly similar to strategies known from established corpora. The intuitive response to problems posed by the data was generally unanimous across the different teams dealing with different CMC genres, for example, regarding the definition and delimitation of textual units, as well as the handling of linguistic features and typographical conventions that are not encountered in more traditional media. All of the solutions offered in this paper aim at facilitating the conversion of original CMC data into text-only files which can be searched with the usual concordance programmes. The tags proposed are straightforward and easy to implement in any type of digital discourse, allowing other datasets to be tagged along the same lines. Regarding the *DMC* itself, the design and markup opted for in the preliminary version will allow the corpus to expand and include further genres, and further languages, as the project continues.

References

Beißwenger, Michael and Angelika Storrer. 2008. Corpora of computer-mediated communication. In Anke Lüdeling and Merja Kytö (eds.), *Corpus linguistics: an international handbook. Volume 1*. Berlin: Mouton de Gruyter, 292–308.

Capraro, Robert M. and Scott W. Slough (eds.). 2009. *Project-based learning: an integrated science, technology, engineering, and mathematics (STEM) approach*. Rotterdam: Sense.

Crystal, David. 2003. *The Cambridge encyclopedia of the English language*. Second edition. Cambridge: Cambridge University Press.

Crystal, David. 2004. *A glossary of netspeak and textspeak*. Edinburgh: Edinburgh University Press.

Crystal, David. 2006. *Language and the Internet*. Second edition. Cambridge: Cambridge University Press.

Crystal, David. 2010. The changing nature of text: a linguistic perspective. In Wido van Peursen, Ernst D. Thoutenhoofd and Adriaan van der Weel (eds.), *Text comparison and digital creativity*. Leiden: Brill, 229–251.

Crystal, David. 2011. 'O brave new world, that has such corpora in it!' New trends and traditions on the Internet. Plenary paper to ICAME 32: Trends and Traditions in English Corpus Linguistics. Oslo, June.

Facebook. 2013a. Newsroom: Key Facts. *Facebook*. Webpage. <http://newsroom.fb.com/Key-Facts> (9th July 2013).

Facebook. 2013b. Information. *Facebook*. Webpage. <http://www.facebook.com/ facebook?v=info> (9th July 2013).

Ferrara, Kathleen, Hans Brunner and Greg Whittemore. 1991. Interactive written discourse as an emergent register. *Written Communication* 8/1: 8–34.

Herring, Susan C. 2002. Computer-mediated communication on the Internet. *Annual Review of Information Science and Technology* 36: 109–168.

Herring, Susan C. 2007. A faceted classification scheme for computer-mediated discourse. *Language@Internet* 4. Article 1. <http://www.languageatinternet. org/articles/2007/761> (12/06/2013).

Owen, Paul and Christopher Wright. 2009. Our top 10 funniest YouTube comments – what are yours? Blog posting. *The Guardian* "Technology Blog", 3 November 2009. <http://www.guardian.co.uk/technology/blog/2009/nov/ 03/youtube-funniest-comments> (9th July 2013).

Peterson, Eric E. 2011. How conversational are weblogs? *Language@Internet* 8. Article 8.

Siebenhaar, Beat. 2006. Code choice and code-switching in Swiss-German Internet Relay Chat rooms. *Journal of Sociolinguistics* 10/4: 481–506.

Stoller, Fredricka L. 2002. Project work: a means to promote language and content. In Jack C. Richards and Willy A. Renandya (eds.), *Methodology in language teaching: an anthology of current practice*. Cambridge: Cambridge University Press, 107–119.

Wrigley, Heide Spruck. 1998. Knowledge in action: the promise of project-based learning. *Focus on Basics* 2/D: 13–18.

Yates, Simeon Y. 2001. Researching Internet interaction: sociolinguistic and corpus analysis. In Margaret Wetherell, Simeon Yates and Stephanie Taylor (eds.), *Discourse as data: a guide for analysis*. London: SAGE, 93–146.

Yus Ramos, Francisco. 2011. *Cyberpragmatics: Internet-mediated communication in context*. Amsterdam: John Benjamins.