2 abbrevi8 or not 2 abbrevi8: A contrastive analysis of different shortening strategies in English and German text messages

Markus Bieswanger
Johann Wolfgang Goethe-University, Frankfurt, Germany

1. Introduction

Text messaging, also called Short Message Service (SMS), refers to the ability to send and receive text messages on cellular telephones. A standard text message can be up to 160 characters in length including spaces when Latin alphabets are used and up to 70 characters when non-Latin alphabets are used. Since its commercial launch in 1995, text messaging, also known as texting, has experienced enormous growth. Today it is supported by virtually all cellular phones and networks around the world, and facilitates communication in a wide variety of languages:

> Indeed, SMS has become a global phenomenon, with billions of text messages sent worldwide every week. It is estimated that a worldwide total of 1 trillion text messages were sent in 2005. (GSMWorld, 2006)

Texting is so popular that it has given the English language the verb *text* that was included in the OED in 2004:

> **text**, *v.*
> *trans Telecomm.* To send (a text message) to a person, mobile phone, etc.; to send a text message to. *Also intr.* to communicate by sending text messages. Cf. TEXT MESSAGE *v.*

From the very beginning, laypeople, journalists and many linguists have claimed that text messages are characterized by a heavy use of shortening strategies. For example, Sutherland emphasizes in *The Guardian* on November 11, 2002, that "[a]bbreviation is the essence of texting" and Crystal (2001) makes a similar claim in his much-quoted book *Language and the Internet*:

> "The challenge of the small screen size and its limited character space (about 160 characters), as well as the small keypad, has motivated the evolution of an even more abbreviated language than emerged in chatgroups and virtual worlds." (p. 229)

These and similar claims about the linguistic properties of text messages have been for the most part based on intuition and anecdotal evidence from English text messages rather than linguistic analysis of empirical data. According to this view, shortenings are presented to be the one major characteristic of text messaging that is assumed to be technologically determined by the limited number of permitted characters and the cumbersome input via the small cellular phone keypad. (Over-) Generalizations of this kind about so-called "texting language", or what Crystal initially refers to as a part of "Netspeak"[1] (2001, p. 18) and later calls "Textspeak" (Crystal, 2004), – frequently made in English-speaking contexts – do not necessarily hold true when applied to messages written in languages other than English. In fact, when we analyze messages from different languages, the occurrence of shortening strategies differs in terms of frequency and structure. This paper is based on an exploratory contrastive study that shows that the shortening strategies in text messages differ fundamentally by respective preferences for certain kinds of shortenings as well as the average number of shortenings per message between a corpus of English text messages and a collection of text messages written in German.[2]

2. Hypothesis and Research Questions

Although not strictly "computer-mediated", text messaging as a form of technologically-mediated and text-based communication is now frequently counted among the modes of computer-mediated communication (CMC) and computer-mediated discourse (CMD) (Thurlow, 2003). As all forms of CMC, text messaging is characterized by specific technological properties, which are commonly referred to as "medium variables" (Herring, 2001, p. 614) in CMC research. These variables were originally proposed to describe the technological properties of communication by computer networks and have to be adapted to fit the mode of texting. The most important medium variables that are shared by text messages in different languages are:

1. Like most CMC, text messaging is text-based.
2. Texting, like e-mail, is an asynchronous mode of CMC. This means that the sender and the receiver do not have to be present at their machines, such as their computers or in this case their cellular phones, at the same time in order to send or receive messages.
3. The transmission process is one-way, i.e. there is no possibility of simultaneous feedback as in telephone voice conversations or face-to-face interaction.
4. Text messages using the Latin alphabet cannot exceed 160 characters including spaces.[3]
5. Text messages have to be typed on the small keypad of a cellular phone, using the limited number of keys available.[4]

[1] It is particularly interesting to note that Crystal (2001) explicitly emphasizes the wider scope and the advantage of his term "Netspeak" over terms such as "Netlish", which he says "is plainly derived from 'English', and is of decreasing usefulness as the Net becomes more multilingual" (p. 17).

[2] According to Danet/Herring (2003), "[t]o date, the research literature in English on computer-mediated communication has focused almost exclusively on emergent practices in English, neglecting developments within populations communicating online in other languages." Contrastive studies on the language of texting are virtually non-existent, with the exception of Schlobinsky/Watanabe (2003).

[3] Some cellular phones and networks now allow their users to send text messages that are longer than 160 characters, but these messages are delivered and billed in multiple segments of 160 characters each.

[4] Text messages are occasionally also sent from computers to cellular phones, but the vast majority of text messages are typed using the keypad of cellular phones.

Despite these shared variables, I had the impression, when communicating with native speakers of English and German via text messages in their respective mother tongue, that English and German text messages seem to show different characteristics when it comes to shortenings, namely that there are much fewer shortenings in messages written in German and that there are clearly differing preferences for certain kinds of shortenings in each language. The hypothesis of this study thus is that text messages in English and German are different with respect to shortenings, despite the shared technological variables of German and English texting and general claims about texting such as Crystal's and Sutherland's (see above). In connection with this hypothesis, two research questions will be addressed in this paper: Firstly, does the average number of shortenings differ between messages written in English and German? Secondly, are there preferences for certain types of shortenings in the messages written in the respective languages?

3. Methodology and Data

Two corpora of text messages, i.e. one collection of text messages composed in English as well as a corpus of messages written in German, have been compared to adequately address the research questions posed in this exploratory study.

The English-language corpus used, which can be found at www.netting-it.com, consists of 201 text messages[5] and was compiled as a sample corpus to accompany Shortis' (2001) textbook *The Language of ICT – Information and Computer Technology*. The messages were collected in the United Kingdom around the year 2000. The content of the messages clearly indicates a college context; most messages in the corpus were most likely written by college-age and immediately post-college-age cellular phone users.[6] The senders and receivers were both male and female, but the exact number is not specified.

The analyzed text messages in German have been selected from a corpus of around 1500 text messages that was compiled by students in two German cities in 2001 and is available at http://www.mediensprache.net/archiv/corpora/sms_os_h.pdf[7]. The original corpus is subdivided into age groups and by gender of the text message senders. To ensure comparability, only the 387 messages written by 17 to 30 year-old females and males have been analyzed to best match the social background of the senders of the English-language material.

Shortening strategies in many forms of CMC include both syntactical as well as lexical reductions (Döring, 2002; Hård af Segerstad, 2002). This paper concentrates on lexical reductions that have been organized into six broad categories, as outlined in the following section.

4. Categories of Shortenings

The term "shortening" is used in this paper as a neutral term to cover all forms of lexical shortening strategies. Shortenings in the sense of this study are all lexical

[5] It actually consists of 202 messages, but I have deleted one message by a sender who explicitly mentions that she is too drunk to write a comprehensible text message.

[6] Unfortunately it has not been possible to obtain additional information on the corpus from the author, such as the exact social background of the senders and receivers, as all email requests to the provided contact addresses have been returned undeliverable.

[7] This corpus is the basis of an early study of SMS-language in German (Schlobinki et al., 2001).

forms that are made up by fewer characters than the full form of a word or a combination of words. The following six categories cover all shortenings used in the two corpora and will be defined in more detail below: initialisms, clippings, contractions, letter/number-homophones, phonetic spellings, and word-value characters (the definition of the first two categories to a large extent follows the terminology in López Rúa, 2002).[8]

Initialisms

Initialisms are shortenings that consist of the first letter (or letters) of a combination of more than one word. The subdivision of initialism into acronyms, i.e. initialisms that are pronounced as one word such as *laser* or *NATO*, and alphabetisms, i.e. initialisms that are pronounced letter by letter such as *BBC* or *NHS*, does not play a role in the context of this study. Examples from the corpora are:

(1)	English:	*NY*	New Year
(2)	German:	*HDL*	hab dich lieb ('love you')

Clippings

Clipping refers to all forms of shortening by which parts of a word are deleted. Clipping here is thus not only the deletion of letters at the end of a word, which Cannon (1989, p. 108) calls "traditional clipping", but includes forms that show letter deletion at the front, i.e. initial clipping, letter deletion in the middle, i.e. medial clipping, and letter deletion in different places in the same word, i.e. mixed clipping. All forms that are shorter than the original word and preserve some of the original letters without adding extra letters that do not belong to the original word are thus clippings. Consider this small selection of examples from the corpora:

(3)	English:	*gettin*	getting
		bday	birthday (also in the form *b'day*)
(4)	German:	*Antw*	Antwort ('answer')
		mal	einmal ('one time')

Contractions

Contractions are combinations of two words that lead to a smaller number of characters than the spelling of the two words individually. Contractions are similar to medial clippings in that letters are usually deleted from the middle of the new combination. These are just some of the many examples in the corpora:

(5)	English	*don't*	do not (also as shorter *dont*)
		were	we are[9]
(6)	German	*hab's*	habe es ('have it')
		auf'm	auf dem ('on the')

[8] The usage of these labels has to be explained in some detail, as there is no agreed standard terminology for the description of shortenings, particularly when it comes to the description of shortenings in CMC. For example, Thurlow (2003) calls *lab* representing *laboratory* a shortening and opposes shortenings to clippings such as *hav* for *have* and *cardif* for *Cardiff*, whereas Shortis (2001, p.104) gives *lab* as a typical example for the phenomenon of clipping.

[9] The spelling of this contraction without an apostrophe results in a homographic clash with *were*, the past tense plural form of *be*.

Letter-/Number-Homophones

Letter-/Number-Homophones are among the most salient features of text messaging. Letters and numbers whose pronunciation is identical with words or parts of words are used to replace words or letter sequences. Crystal (2001, p.229) refers to this phenomenon as "rebus-like potential" of letters and numbers. There are no examples of this kind of shortening in the German collection of messages but numerous examples in the English-language corpus:

(7)	English	*b*	be
		c	see
		l8r	later
		2	to, too

Phonetic Spellings

Phonetic spellings in this context are all forms that are shorter than the original word they represent and go back to the pronunciation of the respective word. These spellings are different from clippings in that they contain at least one character that is not part of the standard spelling of the word in question. The following examples from the corpora may help to illustrate this category:

(8)	English	*bin*	been
		nite	night
(9)	German	*leida*	leider ('unfortunately')
		net	nicht ('not')

Word-Value Characters

Word value characters are a special category that is made up of characters or combinations of up to three characters that can stand for whole words but whose pronunciation is not homophonous with a word. Some of these characters could be treated as extreme cases of clipping, but as we are mostly concerned with individual characters representing whole words, these characters are treated separately here. Examples for the characters are:

(10)	English	*x*	kiss
		&	and
(11)	German	*x*	mal ('times' as in *2 times 4 is 8*)
		h	Hannover ('Hanover, Germany') or Uhr ('o'clock')
		FL	Flensburg

5. Overall Frequency of Shortenings

One of the research questions underlying this paper refers to the previously-mentioned claim that texting is generally characterized by a heavy use of shortenings and whether this claim holds true for languages other than English.

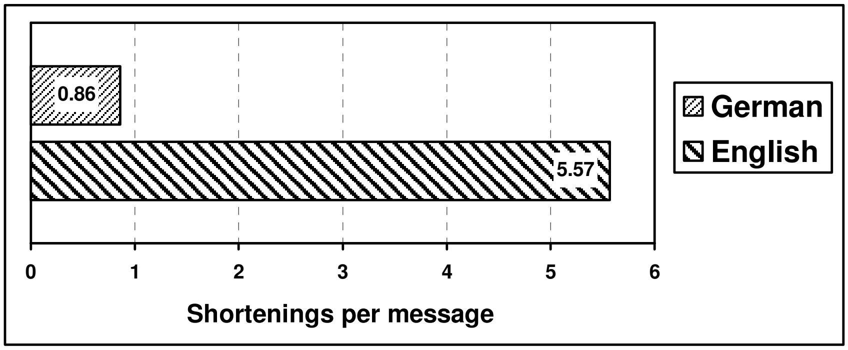
The average length of the text messages in the English-language corpus is roughly 91 characters per message, while the average length of the messages written in German is 95 characters per message. The similarity of average length allows us to compare the overall frequency of shortenings in the two corpora using figures for the average shortening tokens per message. Capitalization will be ignored, as many cellular phone users use only capitals when texting and many phones capitalize all first letters after certain punctuation marks. Consider the following message from the English corpus:

(12) Thanx 4 the time we've spent 2geva, its bin mint! Ur my Baby
and all I want is u!xxxx

This message contains the following shortening types: *thanx*, *4*, *we've*, *2geva*, *its*, *bin*, *u*, *r* and *x* representing kisses. There are two tokens of the type *u* and four tokens of the type *x*. All other types occur only once. The number of shortening types is thus nine, while the number of tokens is thirteen.

The overall frequency of shortenings per message in each of the corpora is the overall number of shortening tokens divided by the number of text messages analyzed. A total of 1120 tokens are contained in the 201 messages of the English corpus, i.e. on an average there are 5.57 shortenings in each message. In the German corpus, a total number of 334 tokens can be found in 387 messages, i.e. the average number of shortenings per message is 0.86. Shortenings are thus more than six times as frequent in the analyzed English text messages when compared with text messages in the German corpus, as illustrated in the following chart:

(13) Shortenings per text message (total)



The comparison of the English and the German corpora with respect to the overall frequency of shortenings in text messages shows impressively that shortenings are much more frequent in English text messages than in German messages sent by a similar social group. The fairly small number of shortenings found in the German corpus corresponds to the results of a study on shortening in German text messages conducted by Döring (2002), who analyzed 1000 text messages written by 124 German students and reports that the number of lexical reductions was surprisingly small. In a cross-medial comparison, Döring (2002) even found the number of shortenings in German text messages to be lower than in German-language newspapers.

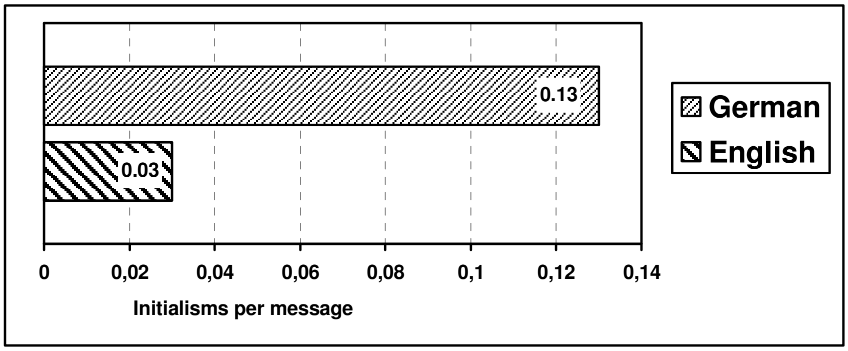
6. Frequency by Shortening Category

After considering the overall frequency of shortenings in the text messages of the German and English corpora, we will now identify language-specific preferences for certain kinds of shortenings by comparing the frequency of these groups of shortenings in the data.

Frequency of Initialisms

The English corpus contains only five different types of initialisms with a total token number of 6, whereas the German data contains 25 types and 50 tokens of initialisms. This means that there are on average 0.03 tokens of initialisms per English text message, as opposed to 0.13 initialisms per German message:

(14) Frequency of initialisms per message

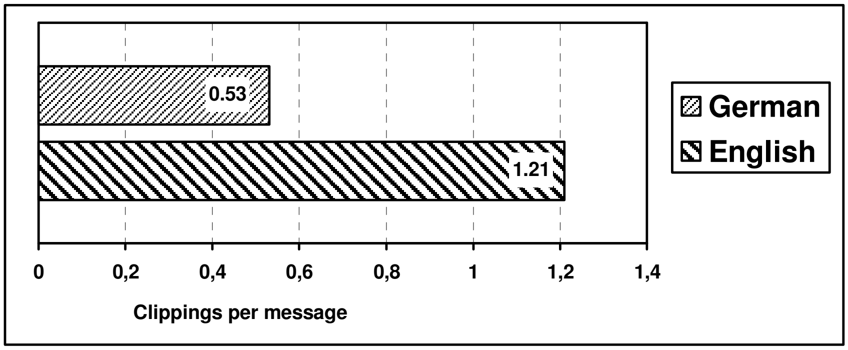


Taking the much higher total number of shortenings in English text messages into account (cf. section 5), the strong preference for initialisms in the German messages of the data appears to be even more marked.

Frequency of Clippings

Clippings are frequently found in both corpora. There are 121 types and 244 tokens in the English corpus and 136 types and 207 tokens in the German corpus. This means that there is an average of 1.21 clippings per message in the analyzed English text messages and 0.53 clippings per German message:

(15) Frequency of clippings per message

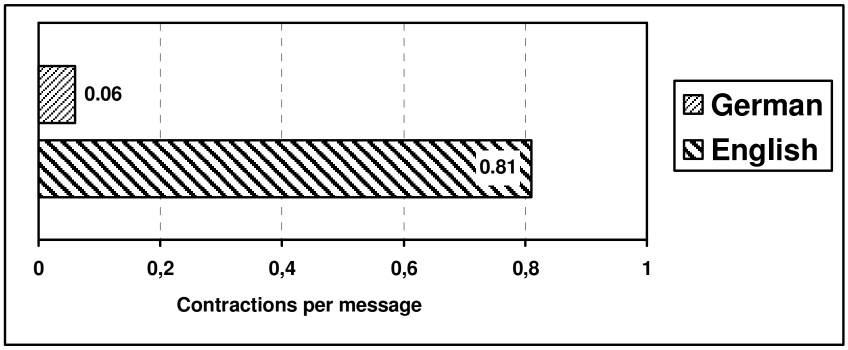


Clippings are obviously more common with the English senders of text messages than with the German writers, but the difference in frequency is still much smaller than the gradient between English and German in overall frequency.

Frequency of Contractions

Contractions are much more frequent in the English data in the analyzed German text messages. There are 57 types and 163 tokens in the English corpus but only 17 types and 21 tokens in the German data. This works out as an average of 0.81 contractions per English message as opposed to only 0.06 contractions per German message:

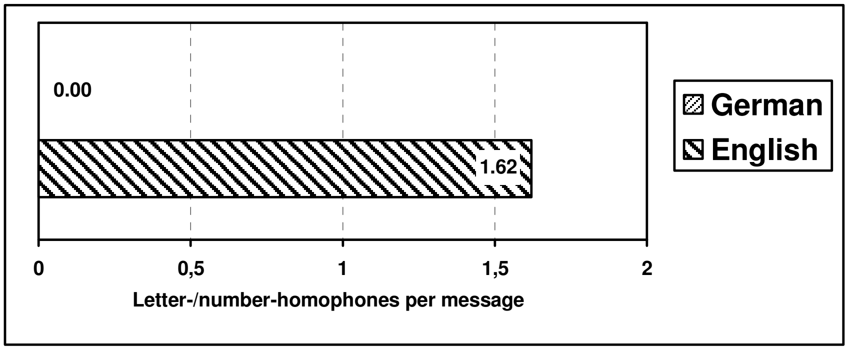
(16) Frequency of contractions per message



Frequency of Letter-/Number-Homophones

Letter-/Number-Homophones are in widespread use in English and can frequently be found in English text messages, as opposed to German where letter-/number-homophones play no role. This is reflected by the data, as there are 326 tokens of letter-/number-homophones in the English corpus, i.e. an average of 1.62 letter-/number homophone per English message, but no letter-/number-homophones in the German data at all.

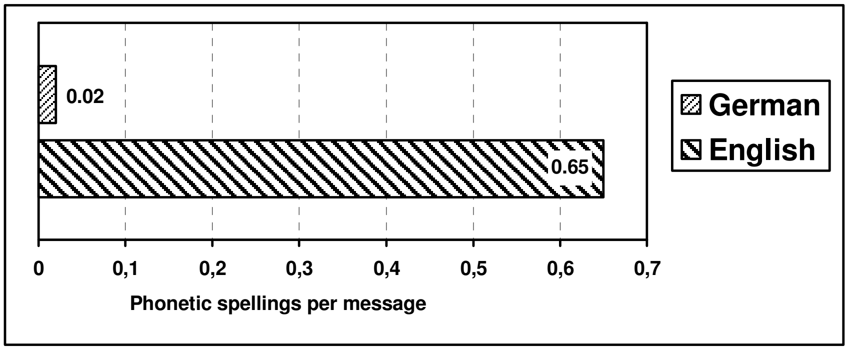
(17) Frequency of letter-/number-homophones per message



Frequency of Phonetic Spellings

Phonetic spellings play only a very minor role in the German, with 6 tokens in 387 messages or 0.02 tokens per message. In the English corpus, on the other hand, phonetic spellings are rather frequent, i.e. there are 131 tokens or 0.65 occurrences per message.

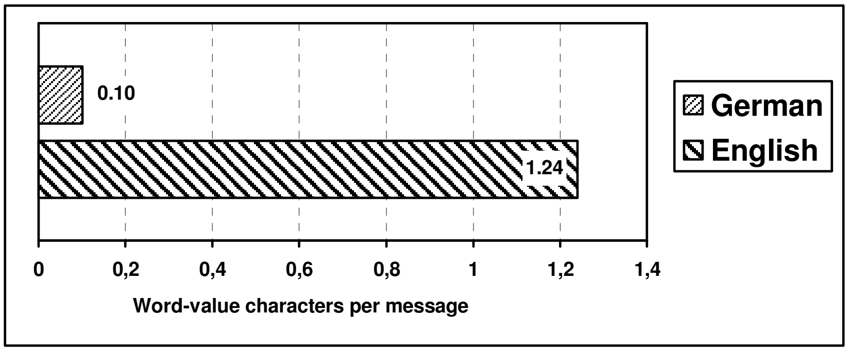
(18) Frequency of phonetic spellings per message



Frequency of Word-Value Characters

Word-value characters require a more detailed analysis, as there are two different aspects of differences between the results from the English and the German corpus and the sole interpretation of token numbers may be misleading. For example, individual letters or combinations of up to three letters are used on German license plates to indicate the city or administrative district where the car is registered. These combinations of characters are frequently employed by senders of text messages in German to talk about the respective cities, for example, *h* for Hanover*, hh* for *Hamburg* or *wob* for *Wolfsburg*. There are 9 types and 18 tokens of this kind of usage in the corpus, plus several other word-value characters making it a total of 22 types and 39 tokens of word-value characters in the German data. The analyzed messages in English contain only four types but 250 tokens, as the type *x* representing 'kiss' occurs 225 times in the data. The average number of occurrence of word-value characters is 1.24 per English message and 0.10 per German message:

(19) Frequency of word-value characters per message



7. Conclusion

Concerning the two research questions formulated in section 2 of this paper, the results of this exploratory study can be summarized as follows: Firstly, the overall frequency of shortenings per text message in the corpus of English text messages is more than six times as high as in the German corpus, namely on average 5.57 shortenings per English message as opposed to only 0.86 messages per German message. The result of the analysis supports the hypothesis that there seem to be considerably fewer shortenings in text messages written in German than in text messages written in English. Large scale studies of structurally and socially identical parallel corpora of English and German text messages are needed to make a more general claim.

Secondly, the contrastive analysis of the individual categories of shortenings and the comparison of the respective frequencies in the corpus data suggest that there seem to be pronounced differences with respect to preferences for certain kinds of shortenings in English and German text messages. Initialisms are much more frequent in the messages from the German corpus, clippings are fairly frequent in both corpora, and contractions and phonetic spellings are much more frequent in the messages in the English corpus. An average of 1.64 letter-/number-homophones per message in the English corpus make this category of shortenings the most frequent individual kind of shortening. This observation is particularly interesting, as there are no letter-/number-homophones in the German corpus at all, as the German language does not provide the same potential in this category as English. Word-value characters show a remarkable distribution in the corpora in that the token total of this category of shortenings is much higher in the English corpus, but the number of different types is more than 5 times higher in the German data.

The average number of 5.57 shortenings per English text message of an average length of 91 characters raise questions whether claims such as the ones made by Sutherland (2002, cf. section 1) and Crystal (2002, cf. section 1), which define shortenings as the sole or major defining characteristic of text messages, are really valid for text messages written in English. Inter-medial and inter-modal studies are needed to clarify this issue. It is, however, certain that the statements definitely do not hold true cross-linguistically when being applied to text messages written in other languages, in this case in German, as the overall number of shortenings in the analyzed German text messages is extremely low.

A further conclusion that goes beyond the original research questions of this paper can be draw from the results of this study, even though the issue cannot be discussed in detail within the scope of this paper: The average message lengths of around 95 characters in the German corpus and 91 characters in the English corpus together with the significant cross-linguistic differences in overall frequency of shortenings cast serious doubt on the claim that the technologically "limited character space" (Crystal 2001, p. 229; cf. section 1) is among the main motivations for the alleged need to use heavily abbreviated language in text messages. If this were really the case, the language variable would have to be essentially irrelevant, as the limit of 160 characters and the other technological variables (cf. section 2) apply to all text messages using the Latin alphabet. Thurlow (2003, section 3.1) makes a similar observation with respect to message length and comments that "[w]hile much is made about the technologically imposed need for brevity in SMS, our participants' messages seldom used the space available." Language-specific linguistic factors, such as language structure and the availability of commonly used shortening strategies, and extra-linguistic motivations, such as the desire to appear "witty" by playing with language, seem to play an important role for the motivation to use shortenings.

Thurlow (2006) emphasizes "that generalizations about CMC are inherently problematic, conflating as they do important differences in the specific affordances and communicative practices of different technologies." This paper has shown that (over-) generalizations about the individual modes of technologically-mediated communication, in this case the language of texting, are also dangerous, especially when they are based on an essentially monolingual perspective. Much more data-based systematic and contrastive research will be necessary before we can make any qualified claims about language use in texting, particularly at a multilingual level.

References

Cannon, G. (1989). Abbreviations and Acronyms in English Word-Formation. *American Speech*, *64 (2)*, 99-127.

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

Crystal D. (2004). *A Glossary of Netspeak and Textspeak.* Edinburgh: Edinburgh University Press.

Danet, B. & S. C. Herring. (2003). Introduction: The Multilingual Internet. *Journal of Computer-Mediated Communication, 9 (1)*. Retrieved July 1, 2006, from http://jcmc.indiana.edu/vol9/issue1/intro.html

Döring, N. (2002). "Kurzm. wird gesendet" – Abkürzungen und Akronyme in der SMS-Kommunikation. *Muttersprache - Vierteljahresschrift für deutsche Sprache, 112 (2),* 97-114.

GSMWorld. (2006). Messaging. Retrieved July 1, 2006, from http://www.gsmworld.com/services/messaging.shtml

Hård af Segerstad, Y. (2002). *Use and Adaptation of Written Language to the Conditions of Computer-Mediated Communication.* Göteborg: Göteborg University.

Herring, S. C. (2001). Computer-mediated discourse. In D. Schiffrin, D. Tannen, & H. Hamilton (Eds.), *The Handbook of Discourse Analysis* (pp. 612-634). Oxford: Blackwell.

López Rúa, P. (2002). On the structure of acronyms and neighboring categories: a prototype-based account. *English Language and Linguistics 6 (1)*, 31-60.

Shortis, T. (2001). *The Language of ICT – Information and Communication Technology*. London/New York: Routledge.

Sutherland, J. (2002, November 11). Cn u txt? *The Guardian*. Retrieved July 1, 2006, from http://www.guardian.co.uk/mobile/article/0,2763,837709,00.html

Schlobinski et al. (2001). Simsen. Eine Pilotstudie zu sprachlichen und kommunikativen Aspekten der SMS-Kommunikation. *Networx 22*. Retrieved July 1, 2006, from http://www.mediensprache.net/networx/networx-22.pdf

Schlobinski, P. & M. Watanabe. (2003). SMS-Kommunikation – Deutsch/Japanisch kontrastiv. Eine explorative Studie. *Networx*. Retrieved July 1, 2006, from http://www.mediensprache.net/networx/networx-31.pdf

Thurlow, C. (2003). Generation Txt? The sociolinguistics of young people's text-messaging. *Discourse Analysis Online, 1* (1). Retrieved July 1, 2006, from http://www.shu.ac.uk/daol/articles/v1/n1/a3/thurlow2002003-paper.html

Thurlow, C. (2006). From statistical panic to moral panic: The metadiscursive construction and popular exaggeration of new media language in the print media. *Journal of Computer-Mediated Communication, 11 (3)*. Retrieved July 1, 2006, from http://jcmc.indiana.edu/vol11/issue3/thurlow.html

Corpora:

English:	http://www.netting-it.com
German:	http://www.mediensprache.net/archiv/corpora/sms_os_h.pdf

Markus Bieswanger
Department of Linguistics at the IEAS
Johann Wolfgang Goethe-University Frankfurt

Grüneburgplatz 1 – Fach 142
60323 Frankfurt
Germany
m.bieswanger@em.uni-frankfurt.de