Computers in Human Behavior 34 (2014) 187–193

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

Letter repetitions in computer-mediated communication: A unique link between spoken and online language

Yoram M. Kalman [a,*], Darren Gergle [b]

[a] The Open University of Israel, 1 University Road, Ra'anana 43107, Israel
[b] Northwestern University, 2240 Campus Drive, Evanston, IL 60208, USA

ARTICLE INFO

Article history:
Available online 25 February 2014

Keywords:
CMC cues
Nonverbal cues
Computer-mediated communication
Letter repetitions

ABSTRACT

Computer-mediated communication (CMC) affords many CMC cues which augment the verbal content of the message: all uppercase letters, asterisks, emoticons, punctuation marks, chronemics (time-related messages) and letter repetitions, to name a few. Letter repetitions are unique CMC cues in that they appear to be a written emulation of a spoken paralinguistic cue – phoneme extension. In this study we explore letter repetitions as a CMC cue, with specific emphasis on elucidating the link between them and spoken nonverbal cues. The letter repetitions are studied in the Enron Corpus, a large ecologically valid collection (∼500,000) of e-mail messages sent by and to employees of the Enron Corporation. We conclude that letter repetitions in the corpus often, but not always, emulate spoken nonverbal cues. This conclusion is examined in a longitudinal analysis that demonstrates the dynamic nature of this cue, and suggests that the usage of letter repetitions is increasing over time, while the link to spoken language is diminishing.

1. Introduction

One of the tools used to convey important social and relational information in computer-mediated communication (CMC) are CMC cues.[1] The information the cues convey cannot be extracted from the lexical or literal meaning of the words that comprise the message, and their creation and interpretation are context dependent and complex. These characteristics of CMC cues are reminiscent of the characteristics of nonverbal cues in traditional communication (Burgoon & Hoobler, 2002). These traditional cues have been defined as "those behaviors that could reasonably function as messages within a given speech community. More specifically, it includes those behaviors other than words themselves that form a socially shared coding system" (p. 244). In this paper, we use the term CMC cues as an analog to traditional nonverbal cues, and define CMC cues as *those modifications of a CMC message that, within a socially shared coding system, modify the meaning of the message while preserving the words of the message and their sequence.*

This paper focuses on elucidating the mechanism by which one category of CMC cues, letter repetitions, are used to enrich online language. We begin the introduction with a brief review of the controversy over the richness of online language, and show that although the emerging consensus is that CMC is capable of conveying social and relational information, our understanding of the mechanisms through which this capacity is achieved is inadequate. We then focus on elucidating some of these mechanisms in letter repetitions through an in depth analysis of a large corpus of CMC messages.

Over the past two decades, there has been a great deal of debate in the literature about the richness of text-based computer-mediated communication (CMC). Media richness theory labeled CMC as poor in relation to other media such as face-to-face or phone communication (Daft & Lengel, 1986), and the cues filtered out model emphasized the impoverishment of CMC given its reduced social context cues (Sproull & Kiesler, 1986). Later work tried to explore the impact media leanness has on the outcomes of group decision making (Baltes, Dickson, Sherman, Bauer, & LaGanke, 2002; Dennis & Kinney, 1998), on online collaboration (Kerr & Murthy, 2009), in very large groups (Lowry, Romano, Jenkins, & Guthrie, 2009), and more (e.g. Otondo, Van Scotter, Allen, & Palvia, 2008; Sia, Tan, & Wei, 2002). The results suggest that the early theories could not account for the mounting evidence that CMC is being used extensively and effectively in contexts requiring subtle interpersonal and socially-oriented communication. More contemporary frameworks such as social information processing (SIP) and social identity/deindividuation (SIDE) theory (Walther,

* Corresponding author. Tel.: +972 9 778 0941; fax: +972 9 778 2668.
E-mail addresses: yoramka@openu.ac.il (Y.M. Kalman), dgergle@northwestern.edu (D. Gergle).

[1] The term CMC cues was suggested by Prof. Joe Walther in a personal conversation.

http://dx.doi.org/10.1016/j.chb.2014.01.047
0747-5632/© 2014 Elsevier Ltd. All rights reserved.

2011; Walther & Parks, 2002) explore the conditions under which CMC is as effective as traditional modes of communication, or even more effective. Both SIP and SIDE acknowledge that CMC does not transmit the same nonverbal cues that traditional spoken conversation does. Both also emphasize the importance of the cues which are transmitted in CMC. SIP puts special emphasis on chronemic cues and the importance of time in online communication (Walther, 2002). SIDE emphasizes paralanguage, which includes alternative usage of characters in the written message such as capitalization, spelling, and punctuation marks (e.g. Lea & Spears, 1992). We review the evidence for the existence of CMC cues, their prevalence, and their usage, as well as the relatively scant research on the mechanisms that enable CMC to convey these socio-emotional cues. Following the review, we focus on one category of cues, letter repetitions, and explore their link to spoken nonverbal cues. We demonstrate the strength of this link in a large corpus of email messages from the late 20th century. In our discussion of these findings we present evidence that the usage of this CMC cue is dynamic, and that as its usage increases over time, the link to spoken language diminishes.

1.1. The cues we use online

In this section we review the cues used in CMC, starting with those that received more extensive attention in past research, namely chronemic cues and emoticons, and continuing with those that have not been studied as extensively. We conclude with a proposed definition for all CMC cues.

One category of cues that has been extensively studied with respect to its role in social communication is chronemics. Chronemics refers to time-related messages and the ways in which the temporal aspects of messaging influence communication. The pioneering experimental study of chronemic nonverbal cues in e-mail by Walther and Tidwell (1995) showed that response latency, as well as the time of day a message is sent, can influence one's perception of the communicator. They also demonstrated that these chronemic cues are context sensitive and can interact with message valence. Later studies of CMC chronemics further demonstrated how chronemic cues can influence the ways in which communicators perceive and make attributions about the social and interpersonal characteristics of those with whom they are communicating (Döring & Pöschl, 2008; Kalman & Rafaeli, 2011; Sheldon, Thomas-Hunt, & Proell, 2006).

Another category of cues that has received extensive attention is emoticons. Emoticons are graphical icons that express emotion, through the representation of a human face. They have been shown, under some conditions, to impact message interpretation (e.g. Derks, Bos, & von Grumbkow, 2007; Walther & D'Addario, 2001). Not unlike nonverbal cues in traditional communication, emoticons are employed in a highly context sensitive manner (Huffaker & Calvert, 2005; Wolf, 2000).

While chronemic cues and emoticons are the two most extensively investigated cues in the literature, there exist a large number of other CMC cues. One of the earliest experimental manipulations of these cues is described in a paper by Lea and Spears (1992). They describe two studies which explore the role of what they labeled as paralinguistic cues in CMC. In the first study, the messages either included or did not include (1) a spelling error in two words in the message; (2) two mistyped words in the message in which the sequence of a pair of letters was reversed; and (3) exclamation marks that were added to the end of one sentence and ellipses at the end of another. The results showed that minor changes in the paralinguistic content of the messages had a significant influence on the impression subjects formed of the anonymous authors of the messages. In the second study, the investigators collected transcripts of online discussions that took place between partners who were either individuated or de-individuated, and who were placed under high or low group salience conditions. The transcripts were analyzed for a series of paralinguistic cues (ellipses, inverted commas, question marks and exclamation marks, as well as sequences of symbols). The results showed significant correlation between paralanguage use and perceived personal attributes. For example, in a high group salience condition there was a strong positive correlation between the use of these paralinguistic cues and measures such as warmth, dominance, liking and responsibility. In the low group salience condition the correlation was either weakened or reversed. These studies lend support to the notion that paralanguage can be a conduit of social information in CMC. In a later study, Postmes and colleagues (Postmes, Spears, & Lea, 2000), looked at the distribution of the same cues, as well as additional cues, in online groups that formed among students taking an academic course. The other cues included nonconventional spelling, deliberately distorted spelling, use of foreign language, capital letter "shouting", message length and chronemic aspects of the communication such as time of day and communication frequency. They show the gradual formation of diverse CMC styles in the different groups, styles which are defined by some of the CMC cues, but not by other cues. This is further evidence for the social meaning of CMC cues. Additional evidence for the role of CMC cues other than chronemics and emoticons in social communication comes from a study of short-message system (SMS) messages posted to a public interactive TV website (Herring & Zelenkauskaite, 2009). An analysis of the properties of 160-character SMS messages posted to the website showed that every message had 8–9 nonstandard typographic features, and that a gender difference exists in relation to the usage of this nonstandard typography: women used more repeated punctuation and more insertions in their messages. The authors conclude that "the resources of written language are employed variably to communicate social meanings that are traditionally conveyed through speech" (p. 27).

While these latter studies begin to expand the notion of CMC cues beyond that of chronemic cues and emoticons, there still exist a large number of relatively unexplored cues in text-based CMC. In the next paragraph we describe some of the key studies that attempted to identify and classify text-based CMC cues.

One of the earliest studies of the wide range of CMC cues is Carey's (1980) work on paralanguage in CMC. Carey identified five categories of cues which he designated as vocal spelling (e.g. "biznis" and "weeeeel"); lexical surrogates and vocal surrogates (e.g. "I like the idea, but then again, it was mine (she said blushingly)" and "hmmm", respectively); spatial arrays which include letters arranged to make a picture, as well as tools such as extra spaces between words to indicate pause or set off a word or a phrase; manipulation of grammatical markers (e.g. multiple exclamation marks or words written in capital letters); and, minus features which is the absence of certain features in the text. This last cue lends a tone to the message such as in the case where no special attention has been given to correcting spelling errors. Another brief exploration of the strategies used to enhance and enrich the written word is by Spitzer (1986) who described a host of typographical devices or "gimmicks", such as usage of capital letters, asterisks, blank spaces, or character repetitions, as well as combinations of these devices. He describes how these are used for emphasis, to show anger, express humor, etc. The next extensive exploration into cues in CMC was by Blackman (1990). This work identified 22 types of nonverbal surrogates. These were divided into seven categories: Kinesic surrogates (kinesic descriptions such as <grin>, kinesic pictographs such as:-), and self pointing such as this arrow pointing at the source's name <===); vocalic surrogates (multiple punctuation marks, all-caps, asterisk bracketing, extended letter repetition, spaces between letters, run-together words, ellipsis, blank spaces in line, vocal characterizations such as (cough), vocal segregates such as er that are used to fill pauses, and interjections such as oops); haptic surrogates (touch descriptions such as KISS and haptic pictographs such as xoxoxo for kisses and hugs); physical appearance surrogates (appearance descriptions and handle pictographs such as (spider/ \ o/\)); artifact surrogates (object displays which occur when a user mentions owning or using some object or substance); action surrogates (action descriptions and sound effects); and miscellaneous (conventional symbols such as $ or #). This study carefully analyzed the frequency of these cues in synchronous and asynchronous CompuServe forums. It reported a rate of about 180 nonverbal surrogates per thousand words in one of the synchronous messaging modes, about 50 per thousand in another synchronous mode, and about 20 per thousand in a third asynchronous mode. Finally, (Riordan & Kreuz, 2010) studied these cues in several contemporary online corpora and identified a frequency of 0.19–0.98% between the different corpora. An automated Linguistic Inquiry and Word Count (LIWC) analysis (Tausczik & Pennebaker, 2010) of one third of the cue laden words revealed that the two largest categories these words fall under are words of affect and words indicating cognitive mechanisms.

Unlike chronemics and emoticons, which have been defined and are studied carefully in various media and contexts, the dozens of other cue categories in the aforementioned studies (Blackman, 1990; Carey, 1980; Riordan & Kreuz, 2010; Spitzer, 1986) have received far less attention. This lack of attention is not surprising, given the resource demanding methodologies required for these studies: careful reading of messages, manual classification of a large number of cues (e.g. Blackman, 1990), and individual interpretation of the meaning of these cues (e.g. Crystal, 2001). Given that these cues are often subtle, highly variable, and that their relative frequency is often low, these studies rarely measured the distributions and identified regularities in the data that could elucidate the possible mechanisms that allow these cues to convey the socio-emotional information.

In this study, we close this gap using methods that enhance manual coding through the power of automated search. This allows us to focus on one specific cue, and explore its usage in an extensive dataset of e-mails. We explore the usage of letter repetitions. This cue has been described in several of the previous descriptive studies reviewed above. It has also been included, aggregated with other cues, in several SIDE-oriented experimental studies which proved the ability of such cues to convey social and relational information. Nevertheless, none of these descriptive or experimental studies attempted to elucidate the principles by which this cue operates in CMC. In this study we aim to collect a sufficiently large and diverse sample of letter repetitions and use it to ask specific research questions about general principles by which these repetitions act as cues in CMC.

1.2. Research questions

In this study we examine the role of letter repetitions in e-mail messages. Like many other CMC cues besides chronemics and emoticons, letter repetitions have not been systematically studied, and their usage is not well understood. Given the uniqueness of letter repetitions and their potential link to spoken nonverbal cues, the goal of the study is to understand the usage of letter repetitions in an ecologically valid, large-scale sample, and to elucidate possible mechanisms for the way CMC cues operate.

The study was conducted in the context of a single dataset (the Enron Corpus, see below) which was extensive and comprehensive, and it employed a search tool that was specifically developed to understand letter repetitions in this corpus. The research question was:

1.3. How are letter repetitions used in email communication?

This general question was split into several subsidiary questions. The first of these is primarily concerned with understanding the link between letter repetitions and spoken communication. Letter repetitions could be interpreted as emulating an extension (repetition) of the phoneme encoded by the repeated letter. If this is so, then it should be possible to vocally articulate the extended phoneme. By answering the question *whether the letter repetitions are articulable or not*, we gain insight into the question whether the letter repetition might be used to convey the equivalent of the spoken paralinguistic cue of extending a phoneme. For more details on how this classification was made, see the Method section.

The second subsidiary question was the result of an anecdotal examination of a small sample of letter repetitions from a collection of emails. The examination revealed that many of the repetitions appeared in onomatopoeic words (e.g. boooo or hmmm). This led to two questions, one specific to onomatopoeic words, and another, more general, about parts of speech: *To what extent are the words that contain letter repetitions onomatopoeic?* And, *which parts of speech do the words that include the letter repetitions belong to?* Answering these questions could provide some evidence as to the function that repetitions play, and provide some insight into repetitions' relation to cues in spoken language.

Further, we asked whether letter repetitions are a cue used by a small subset of the population of those represented in our corpus (described in detail below), or whether it was more widely used. Since CMC cues are meaningful within a specific social context, if a cue is used by a small subset of the group, it would be important to try and identify this group, as well as to limit our conclusions to that group. Thus, we examined *the amount of letter repetitions in the email messages composed by different users in our corpus.*

2. Method

The corpus that was used in this study is the Enron Corpus. This corpus is based on the email archives of Enron Inc., which were confiscated and published online as a part of the U.S. Federal Energy Regulatory Commission's investigation of the company (Berman, 2003). The original dataset was then processed to accommodate the needs of academic researchers and resulted in a corpus of approximately 500,000 e-mail messages in .txt format (Cohen, 2005). This corpus is one of the few publicly available large-scale datasets that contains naturally occurring and unobtrusively collected e-mail messages covering both professional and interpersonal communication sent to and by employees in a commercial company. Some limitations of the dataset are that the e-mails are relatively old, produced no later than 2002, they originate from a single US-based organization, there are many duplicates and corrupted messages in the corpus, and there are sources of noise such as embedded HTML code or large spans of ASCII characters which represent file attachments. As a result, the published dataset required substantial processing as described in the following sections.

2.1. CorpusCruizer

A proprietary Python-based software tool ("CorpusCruizer") was developed to accommodate the study of repetitions in the Enron Corpus. CorpusCruizer supports the efficient analysis of large-scale text collections. The first function of CorpusCruizer is to support flexible search terms for pattern matching using Python Regular Expressions (Python Community, 2013). This allows the identification of every occurrence in the corpus that matches a particular search pattern (e.g. a word that includes a repetition of exactly three lower case m's). Message headers were not analyzed in this study unless they were included in the body of a message due to forwarding, replying with quotes, etc. The second function of CorpusCruizer is to generate a list of all of the occurrences of a specific sequence of characters, and present them in the context of the original flanking text. This *concordance* allows the efficient export and consequent visualization and manipulation of hundreds of snippets of texts that contextualize the requested sequence of characters. Together, the two functions allow the efficient processing of the hundreds of thousands of files in the Corpus, the production of output that reports character string frequencies, and the production of a concordance that permits in-depth exploration of the usage of specific character repetitions in the context of the e-mail messages in which they appear. In this study, CorpusCruizer was used to identify the occurrence and location of repetitions of the 26 letters of the English alphabet.

2.2. Concordance construction and item classification

The first task was to reduce the thousands of occurrences of letter repetitions in the Enron Corpus to a clean list that captures the way repetitions were used in the e-mail messages. In the first stage, the initial result set returned single lines which represented an occurrence of a given letter repetition in the dataset. Since the same word from the same message could appear more than once due to the existence of duplicate messages in the dataset (or due to the replication when a message was included in the body of a forwarded or a reply-to e-mail) this initial list included *dependent occurrences*. Dependent occurrences are two or more occurrences of a specific string which have been typed only once, but that have later been duplicated. In the second stage, all dependent occurrences of a specific repetition were reduced to a single instance (in order to eliminate duplicates as previously defined). This, in effect, produced a list of what we refer to as *independent occurrences*. In the third stage, for each of the independent lines in the result set, a root word was assigned. The root word is the common form of the word that included the repetition. For example, the root word of the entry pleeeeeeze is the word 'please'. Special effort was made to assign each entry to root words as they are spelled in the Oxford English Dictionary (2009). The fourth stage was a data aggregation step, wherein the concordance was organized alphabetically based on the spelling of the root word and appended to each entry as a count of the number of dependent occurrences. For example, if there are two copies of the email message which includes the repetition, as well as one response email that included the original text of the message, then one independent and three dependent occurrences of the repetition were recorded. The dependent occurrences were only recorded and were not used in the analyses.

The second major task was to classify the items according to four major classifications, in order to answer the research questions. The first classification was whether the repetition was articulable or inarticulabe. Inarticulable repetitions for English words were defined as letter repetitions that were part of a plosive consonant. In spoken English it is not possible to vocalize an extended plosive, since plosive consonants require the creation and rapid release of a complete closure of the airflow in the vocal tract (Clark, Yallop, & Fletcher, 2006). If the same word included both an articulable and an inarticulable phoneme (e.g. lllooooonnnngggg where the repeated velar plosive <g> is preceded by repetitions of articulable phonemes), it was classified as both articulable and inarticulable. In the case of a phoneme created by more than one letter (e.g. ck or sh), the repetition of even just one of the letters was interpreted as an elongation of the sound (e.g. russsssshhhh, rushhhhh and russsssh are all equivalent elongations of the same terminal post-alveolar fricative <ʃ> sound). The second classification described the part of speech of the word, under the expectation that specific parts of speech (e.g. adjectives, interjections) will be overrepresented in words that were emphasized using the cue. The categories included noun, adjective, pronoun, verb, adverb, conjunction, preposition, interjection and *other*. The *other* category was used for cases such as words in other languages, abbreviations, acronyms, words which were not in the dictionary, and entries that comprised more than one word (e.g. gonna). The third classification in the coding scheme was whether or not the repetition was onomatopoetic. Onomatopoeic words are those that imitate sounds such as boom or grrr. Words in languages other than English, abbreviations, acronyms, and entries that comprised more than one word were not classified. Finally, the fourth classification recorded either the name or the email address of the message's sender, so as to ensure that the cue is widely used, and is not specific to a subset of the users represented in the dataset.

A small number of messages were excluded from coding: (1) a story about a stuttering person which appeared in many identical copies within the dataset, and included many letter repetitions that expressed stuttering; (2) repetitions used as a part of ASCII art; (3) acronyms which included repetitions such as BBB (Better Business Bureau), XXX which stands for an unknown quantity, etc.; (4) unknown words, and obvious typos; (5) repetitions of xo for hugs and kisses; and (6) repeated characters used to fill in spaces or as a graphical feature (e.g. separating line).

Human coders were used to classify the instances based on this scheme, and 10% of the entries were double-scored. The inter-rater reliability (Cohen's Kappa) for the classification of articulability, part of speech and onomatopoeic words were an acceptable .79, .86 and .77 respectively.

3. Results

3.1. Frequency and usage of letter repetitions in the Enron Corpus

The full concordance included 815 independent entries, representing a total of 2926 occurrences of letter repetitions. Thus, every independent entry had on average 3.6 dependent entries in the dataset (range: 1–54). These entries were collapsed by root word, resulting in a list of 201 root words. The 16 root words which appeared in the dataset ten or more independent times are listed in Table 1.

Of the 815 entries, the vast majority (767) were classified as articulable. For example: "freeeezing". Only 12 were classified as inarticulable (e.g. "butttttttt"), and 36 were classified as including both articulable and inarticulable repetitions (e.g. "lllooooonnnngggg"). The most prevalent part of speech of the entries was interjection (376). Some examples include "Yipeeee", "Aughhhhhh" and "Pssst". This category was followed by adverbs (204), nouns (86), adjectives (62), verbs (30), pronouns (3) and conjunctions (2). Fifty-two of the entries were classified as *other*. The root words of 177 words were classified as onomatopoeic, for example "Bang! Booommm! Craaash!...", 597 were determined not to be onomatopoeic, and 41 were classified as *other*. More than 450 names and e-mail addresses of putative authors of the messages were recorded. An inspection of the list revealed a diversity of names (male and female), as well as of domain names, a finding that precludes the possibility that the usage of letter repetitions is specific only to Enron, or only to a small subset of the people whose messages appear in the dataset. One Enron user was responsible for 49 entries that included letter repetitions. Nineteen more users (mostly from Enron) were responsible for 5–15 entries each, and hundreds of other authors were responsible for the rest of the entries.

The frequencies of repetitions of various lengths in the five most common root words are reported in Table 2. Note that apparent

Table 1
The root words with at least ten independent occurrences in the Enron Corpus, by number of independent occurrences.

Root word occurrences	Number of independent occurrences	Number of dependent occurrences
so	129	494
hm	84	317
ah	33	85
mm	26	66
oops	25	72
oh	22	60
what's up	18	107
whoo hoo	17	50
ooh	15	57
sh	13	38
no	12	31
too	12	35
ugh	12	27
um	11	41
way	11	40
uh	10	32

discrepancies between Table 2 and Table 1 are a result of Table 1 being based on a manual human analysis of the data, while Table 2 is based on an automated count. Thus, for example the personal name Soo was removed from the count that led to Table 1, but is included in the count that led to Table 2.

4. Discussion

This study explores the frequency and usage of letter repetitions in e-mail communication. We conclude that the findings suggest that letter repetitions often, but not always, emulate spoken nonverbal cues as evidenced by the fact that over 94% of the repetitions classified were found to be articulable, and by the disproportional representation of onomatopoeic words in the list of words with letter repetitions. We provide several examples from the corpus that demonstrate this linkage between spoken communication and CMC.

4.1. Letter repetitions often, but not always, emulate spoken nonverbal cues

An inspection of the occurrences of letter repetitions in the Enron Corpus reveals extensive richness and diversity. An inspection of the examples suggests that often letter repetitions are used to emulate spoken nonverbal cues. Here are a few examples.

Repetitions seem to indicate the stretching of a word, emulating a stretched out morpheme in spoken conversation:

> “I was in an electronics store the other night... Panasonic has 9” Portable DVD player (like your sony) with an 8 h battery... $999.00 US. It is sweeeeeeet”.

(This is a direct quote from the Enron Corpus. Corpus quotes appear verbatim indented and between quotation marks).

Or, more playfully:

> “Whaaaassssupppp”

To denote a change in pitch in:

> “Yeeeeeeeeehaaaw!!!!!!!!!!”

To denote or to fill a pause:

> “Hmmmm, I think you’re right. Looks like the more we can get done tonite, the better.”

Or, to express sounds (paralinguistic alternants (Poyatos, 2002):

> “now that i have a ‘temporary’ plate for the harley....... vvvvrrrrooooommmmm.............vvvvvrrrrooooommmmmm!”

To denote musical intonation (in a parody on the song ‘American Pie’):

> “I never worried on the whole way up
> Buying dot coms from the back of a pickup truck
> But Friday I ran out of luck
> It was the day the NAAAASDAQ died
> I started singin’
> Bye-bye to my piece of the pie”

Or, of a birthday song:

> “Happy birthday to youuuu
> Happy birthday to youuuu
> Happy birthday dear frieeeeeeeeeeennnnnd”

To indicate a loud shout:

> “WOOOOOOOOOOHOOOOOOOOOOOOOOOO, Daddy’s getting a new Blue Wave Bay boat!!!! WOOOHOOOO”

To express human-made sounds:

> “And pfffffff, he is away”

Such as laughter:

> “Heeeeeheeee!”

Or guttural sounds:

> “ugggggghhhh!!! what a complete and utter pr–k!! i am SO annoyed reading”

And other sounds:

> “Bang! Booommm! Craaash!...”

What quantitative evidence do we have that letter repetitions often emulate spoken nonverbal cues? The first piece of evidence is that the repetitions were classified as inarticulable in only 12 of the 815 entries, and that only 36 more of the entries included both an articulable and an inarticulable repetition. This provides support for the suggestion that in most cases the repetitions are an attempt to replicate an elongated phoneme that can be articulated in spoken language. About 17% of the phonemes in conversational English are plosives (Mines, Hanson, & Shoup, 1978), and if repetitions were not related to spoken paralinguistic cues (for example if they were visual emphasis markers), we would not expect such a bias against repetitions which are a part of an inarticulable plosive phoneme. A second piece of evidence in support of the link between traditional nonverbal cues and repetitions in CMC is the finding that the root words of 177 of the 815 entries (over 20%) were onomatopoeic words such as *booom* or *shhhhh*. This over representation of onomatopoeic words which are apparently quite

Table 2
Occurrences of repetitions in the Enron dataset for the five most common root words. Search term in parentheses.

182,403 (so)	438 (soo)	159 (sooo)	164 (soooo)	93 (sooooo)
277 (hm)	139 (hmm)	202 (hmmm)	56 (hmmmm)	10 (hmmmmm)
361 (ah)	21 (ahh)	38 (ahhh)	9 (ahhhh)	3 (ahhhhh)
5151 (mm)	126 (mmm)	26 (mmmm)	11 (mmmmm)	14 (mmmmmm)
567 (oops)	59 (ooops)	3 (oooops)	0 (ooooops)	0 (oooooops)

rare in the English language (Katamba, 1994; Sadler, 1971) is an indication that when users try to replicate an audible sound in written communication, they augment the onomatopoeic word with repetitions that help convey the sound's characteristics. In conclusion, the cited examples from the corpus presented above, as well as the findings that the prevalence of letter repetitions increases in words which convey audible sounds, and decreases in inarticulable syllables, all support the suggestion that letter repetitions often, but not always, emulate spoken nonverbal cues.

4.2. Longitudinal analysis of the usage of letter repetitions

The Enron Corpus represents language produced more than a decade ago. The frequency of letter repetitions in this corpus is relatively low: almost 3000 occurrences in more than 500,000 messages comprising about 930 million characters. A recent study (Riordan & Kreuz, 2010) identified 273 instances of letter repetitions in a corpus of about 11 million characters. This rate is almost an order of magnitude higher. Are letter repetitions more prevalent now than they were in the past?

To answer this question we performed an exploratory analysis of the relative frequency of letter repetitions in several root words that often include letter repetitions: the words please, help and oops. The analysis was performed using Google's "blog search" (http://www.google.com/blogsearch), and it compared the (normalized – see below) frequency of repetitions during the five-year period 1998–2002, with their relative frequency during the period 2008–2012. The data were collected using the "custom range" feature that allows limiting a search to blogs that were posted during a custom date range. In order to normalize the frequencies of each word, the number of occurrences of the words that included repetitions was divided by the number of occurrences of the root word. The repetitions included an addition of 2–6 additional identical letters. The results are described in Table 3. For example, the table shows that there were 50 occurences of pleeease, pleeeease, pleeeeease, pleeeeeease or pleeeeeeese (or, in short, pleee*ase) in blogs during 1998–2002, and 418,000 occurrences of the word please in blogs during the same period. These increased to 105,000 and 166 million respectively, during 2008–2012. The ratio of the pleee∗ase repetitions increased from 0.0120% (50/418,000) to 0.0633% (105,000/166,000,000), an increase of 5.29-fold.

These data strengthen the assumption triggered by the Riordan and Kreuz (2010) paper, that letter repetitions are becoming more prevalent with time. Of the nine repetitions presented in Table 3, five have become much more prevalent (5.29-fold, or 529% increase and more), two increased slightly (by about 50%), and two slightly decreased. It is also interesting to see the high increase in the relative prevalence of pleaseee∗, where the repeated e is silent in spoken communication, and of helppp∗, where the terminal p is inarticulable. This raises the possibility that letter repetitions are a cue that is not only becoming more prevalent, but also that the link between the written cue of letter repetition and the spoken cue of elongating the phoneme, is weakening with time. It is possible to hypothesize that the period during which the Enron Corpus was formed (1990s and early 2000s) represents an earlier period during which the letter repetition cue is still strongly linked to "vocal spelling". The cue proved effective due to its high visibility: it draws attention and stands out on the page, and as it became more prevalent, the importance of the visual cue increased, and the strength of the link to spoken language weakened.

Table 3
Relative frequencies of words with letter repetitions during two time periods.

String	1998–2002		2008–2012		Growth
	Occurrences	Ratio (%)	Occurrences	Ratio (%)	
please	418,000	100	166,000,000	100	
pleee*ase	50	0.0120	105,000	0.0633	5.29
ppp*lease	1	0.0002	603	0.0004	1.52
pleaseee*	29	0.0069	196,000	0.1181	17.02
help	664,000	100	207,000,000	100	
heee*lp	4	0.0006	33,900	0.0164	27.19
helll*p	2	0.0003	14,000	0.0068	22.45
helppp*	4	0.0006	140,000	0.0676	112.27
oops	3760	100	13,000,000	100	
oooo*ps	54	1.4362	122,000	0.9385	0.65
ooppp*s	2	0.0532	2530	0.0195	0.37
oopsss*	1	0.0266	5230	0.0402	1.51

Notes. Ratio is calculated by dividing the number of occurrences of a word that includes repetitions by the number of occurrences of the root word. Growth is calculated by dividing the ratio in the later period by the ratio in the earlier period. An asterisk denotes 0–4 additional repetitions of the letter followed by the asterisk. Occurrences are based on estimated number of results provided by Google Blog Search (http://www.google.com/blogsearch).

These findings suggest future research that is beyond the scope of this study. A detailed longitudinal analysis of the usage of the cue requires analyzing the occurrence of the dozens of possible permutations and variations possible for each root word, as was performed on the Enron Corpus, using CorpusCruiser. For example, the root word please could include repetitions such as pleeeeeze, pleeeaze, pleeeeaseeee, pllllleeeasssse, etc. The initial analysis described here is based on a very small number of examples, and on a small number of occurrences in the early (1998–2002) period that could lead to random fluctuations. This is possibly since older blogs have been removed from the Internet, and are no longer indexed by Google.

What is apparent from this preliminary analysis is that the usage of this CMC cue is dynamic and evolving. Unlike nonverbal cues in spoken communication which might vary between cultures, but which are not known to evolve within a short time span of one decade, it seems that CMC cues are still evolving as text-based CMC spreads to additional segments of the populations, involves additional communication media, and becomes a dominant form of communication, as well as with the passage of time.

4.3. Limitations

The main limitations of the study stem from the methodological choice to quantitatively analyze a very large dataset using a computer algorithm, in an effort to detect patterns. Alternative, more qualitative approaches, analyze significantly smaller corpora, but provide more nuanced and context specific conclusions. For example, Vandergriff's (2013) recent study used a microanalytic approach to study different CMC cues used in college classroom discussions. The pragmatic perspective of that study provides insights on the function of different CMC cues in emotive communication. Another recent example is the interactional sociolinguistic approach used by Darics (2013) to closely explore, in context, specific uses of letter repetitions to convey socio-emotional messages, to evoke auditory cues, etc. A final example by Ong (2011) used conversation analysis to closely examine the role of ellipsis in online chat, demonstrating the role of ellipsis as a CMC cue. Another limitation of the study is that it is descriptive, unlike most SIP and SIDE studies which are more experimental in nature. The combination of experimental and descriptive (qualitative and quantitative) methods provides a fuller picture of the use of CMC cues than any of the methods separately.

5. Conclusion

Our findings on the presence and usage of letter repetitions in CMC reveal a link between this textual cue and paralinguistic cues used in spoken conversation. We also present evidence that when emotionally-laden interjections are used, repetitions are more likely to be employed. These findings are in line with the suggestion that letter repetitions in CMC serve as CMC cues that extend the lexical meaning of the words, add character and richness to the sentences, and allow the fine-tuning and personalization of the message. In addition to the theoretical implications of additional support for SIP theory and SIDE, the study's practical implications are that automated analyses of CMC, for example sentiment analysis, should take into account CMC cues such as letter repetitions (Brody & Diakopoulos, 2011). Moreover, it is suggested that this CMC cue is still evolving. Based on a preliminary longitudinal analysis, we hypothesize that as this cue evolves and becomes more popular and prevalent, the link to spoken paralinguistic cues will grow weaker, and the cue will develop a character that is independent of its origins in spoken language. If this hypothesis is supported by future research, it will provide quantitative empirical evidence for language evolution (Christiansen & Kirby, 2003; Huffaker & Calvert, 2005) in computer-mediated communication.

Acknowledgements

We thank Alberto Gonzalez for his work on the development of CorpusCruizer, Amanda Yentz for her work on the Enron Corpus, and Joe Walther for helpful and inspiring discussions. This work was funded, in part, by National Science Foundation grant #0953943.

References

Baltes, B. B., Dickson, M. W., Sherman, M. P., Bauer, C. C., & LaGanke, J. S. (2002). Computer-mediated communication and group decision making: A meta-analysis. *Organizational Behavior and Human Decision Processes, 87*(1), 156–179. http://dx.doi.org/10.1006/obhd.2001.2961.

Berman, D. K. (2003, 5 October). Online laundry: Government posts Enron's e-mail — amid power-market minutiae, many personal items; 'about Wednesday... ', The Wall Street Journal, p. 1.

Blackman, B. I. (1990). *A naturalistic study of computer-mediated communication: Emergent communication patterns in online electronic messaging systems*. Tallahassee, FL, USA: Florida State University.

Brody, S., & Diakopoulos, N. (2011). Cooooooooooooooollllllllllllll!!!!!!!!!!!!!! using word lengthening to detect sentiment in microblogs. In *Proceedings Empirical Methods in Natural Language Processing (EMNLP)*. Edinburgh, UK.

Burgoon, J. K., & Hoobler, G. D. (2002). Nonverbal signals. In M. Knapp & J. Daly (Eds.), *Handbook of interpersonal communication* (pp. 240–299). Thousand Oaks, CA: Sage.

Carey, J. (1980). Paralanguage in computer mediated communication. In *Proceedings of the 18th annual meeting on Association for computational linguistics*. ACM. doi: 10.3115/981436.981458.

Christiansen, M. H., & Kirby, S. (2003). Language evolution: Consensus and controversies. *Trends in Cognitive Sciences, 7*(7), 300–307. http://dx.doi.org/10.1016/S1364-6613(03)00136-0.

Clark, J., Yallop, C., & Fletcher, J. (2006). *An introduction to phonetics and phonology*. Hoboken, NJ: Wiley-Blackwell.

Cohen, W. W. (2005). Enron email dataset. <http://www.cs.cmu.edu/~enron/>.

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

Daft, R. L., & Lengel, R. H. (1986). Organizational information requirements, media richness and structural design. *Management Science, 32*(5), 554–571. http://dx.doi.org/10.1287/mnsc.32.5.554.

Darics, E. (2013). Non-verbal signalling in digital discourse: The case of letter repetition. *Discourse, Context & Media, 2*(3), 141–148. http://dx.doi.org/10.1016/j.dcm.2013.07.002.

Dennis, A. R., & Kinney, S. T. (1998). Testing media richness theory in the new media: The effects of cues, feedback, and task equivocality. *Information Systems Research, 9*(3), 256–274. http://dx.doi.org/10.1287/isre.9.3.256.

Derks, D., Bos, A. E. R., & von Grumbkow, J. (2007). Emoticons and social interaction on the Internet: The importance of social context. *Computers in Human Behavior, 23*(1), 842–849. http://dx.doi.org/10.1016/j.chb.2004.11.013.

Döring, N., & Pöschl, S. (2008). Nonverbal cues in mobile phone text messages: The effects of chronemics and proxemics. In L. R. & S. W. Campbell (Eds.), *The reconstruction of space and time: Mobile communication practices*. New Brunswick, NJ: Transaction Publishers.

Herring, S. C., & Zelenkauskaite, A. (2009). Symbolic capital in a virtual heterosexual market: Abbreviation and insertion in Italian iTV SMS. *Written Communication, 26*(1), 5–31. http://dx.doi.org/10.1177/0741088308327911.

Huffaker, D. A., & Calvert, S. L. (2005). Gender, identity, and language use in teenage blogs. *Journal of Computer-Mediated Communication, 10*(2). http://dx.doi.org/10.1111/j.1083-6101.2005.tb00238.x.

Kalman, Y. M., & Rafaeli, S. (2011). Online pauses and silence: Chronemic expectancy violations in written computer-mediated communication. *Communication Research, 38*(1), 54–69. http://dx.doi.org/10.1177/0093650210378229.

Katamba, F. (1994). *English words*. Routledge.

Keng Wee Ong, K. (2011). Disagreement, confusion, disapproval, turn elicitation and floor holding: Actions as accomplished by ellipsis marks-only turns and blank turns in quasisynchronous chats. *Discourse Studies, 13*(2), 211–234. http://dx.doi.org/10.1177/1461445610392138.

Kerr, D. S., & Murthy, U. S. (2009). The effectiveness of synchronous computer-mediated communication for solving hidden-profile problems: Further empirical evidence. *Information & Management, 46*(2), 83–89. http://dx.doi.org/10.1016/j.im.2008.12.002.

Lea, M., & Spears, R. (1992). Paralanguage and social perception in computer-mediated communication. *Journal of Organizational Computing, 2*(3–4), 321–341. http://dx.doi.org/10.1080/10919399209540190.

Lowry, P. B., Romano, N. C., Jenkins, J. L., & Guthrie, R. W. (2009). The CMC interactivity model: How interactivity enhances communication quality and process satisfaction in lean-media groups. *Journal of Management Information Systems, 26*(1), 155–196. http://dx.doi.org/10.2753/MIS0742-1222260107.

Mines, M. A., Hanson, B. F., & Shoup, J. E. (1978). Frequency of occurrence of phonemes in conversational English. *Language and Speech, 21*(3), 221–241. http://dx.doi.org/10.1177/002383097802100302.

Otondo, R. F., Van Scotter, J. R., Allen, D. G., & Palvia, P. (2008). The complexity of richness: Media, message, and communication outcomes. *Information & Management, 45*(1), 21–30. http://dx.doi.org/10.1016/j.im.2007.09.003.

Postmes, T., Spears, R., & Lea, M. (2000). The formation of group norms in computer-mediated communication. *Human Communication Research, 26*(3), 341–371. http://dx.doi.org/10.1111/j.1468-2958.2000.tb00761.x.

Poyatos, F. (2002). Nonverbal communication across disciplines: Paralanguage, kinesics, silence, personal and environmental interaction (Vol. 2). Amsterdam: John Benjamins Publishing Company.

Python Community (2013). Regular expression operations—Python v2.7.5 documentation. <http://docs.python.org/2/library/re.html>. Retrieved June 25.06.13.

Riordan, M. A., & Kreuz, R. J. (2010). Cues in computer-mediated communication: A corpus analysis. *Computers in Human Behavior, 26*(6), 1806–1817. http://dx.doi.org/10.1016/j.chb.2010.07.008.

Sadler, J. D. (1971). Onomatopoeia. *The Classical Journal, 67*(2), 174–177. http://www.jstor.org/stable/3296513.

Sheldon, O. J., Thomas-Hunt, M. C., & Proell, C. A. (2006). When timeliness matters: The effect of status on reactions to perceived time delay within distributed collaboration. *Journal of Applied Psychology, 91*(6), 1385–1395. http://dx.doi.org/10.1037/0021-9010.91.6.1385.

Sia, C.-L., Tan, B. C. Y., & Wei, K.-K. (2002). Group polarization and computer-mediated communication: Effects of communication cues, social presence, and anonymity. *Information Systems Research, 13*(1), 70–90. http://dx.doi.org/10.1287/isre.13.1.70.92.

Spitzer, M. (1986). Writing style in computer conferences. *IEEE Transactions on Professional Communication, PC-29*(1), 19–22. doi:10.1109/TPC.1986.6449010.

Sproull, L., & Kiesler, S. (1986). Reducing social context cues: Electronic mail in organizational communication. *Management Science, 32*(11), 1492–1512. http://www.jstor.org/stable/2631506.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology, 29*(1), 24–54. http://dx.doi.org/10.1177/0261927X09351676.

The Oxford English Dictionary (2009, June). *Oxford University Press*. <http://oed.com/>.

Vandergriff, I. (2013). Emotive communication online: A contextual analysis of computer-mediated communication (CMC) cues. *Journal of Pragmatics, 51*, 1–12. http://dx.doi.org/10.1016/j.pragma.2013.02.008.

Walther, J. B. (2011). Theories of computer-mediated communication and interpersonal relations. In M. L. Knapp & J. A. Daly (Eds.). *The SAGE handbook of interpersonal communication* (Vol. 4th, pp. 443–479). Thousand Oaks, CA: Sage.

Walther, J. B., & D'Addario, K. P. (2001). The impacts of emoticons on message interpretation in computer-mediated communication. *Social Science Computer Review, 19*(3), 324–347. http://dx.doi.org/10.1177/089443930101900307.

Walther, J. B. (2002). Time effects in computer-mediated groups: Past, present, and future. In P. Hinds & S. Kiesler (Eds.), *Distributed work* (pp. 235–257). Cambridge, MA: MIT Press.

Walther, J. B., & Parks, M. R. (2002). Cues filtered out, cues filtered in. In M. Knapp & J. Daly (Eds.), *Handbook of interpersonal communication* (pp. 529–563). Thousand Oaks, CA: Sage.

Walther, J. B., & Tidwell, L. C. (1995). Nonverbal cues in computer-mediated communication, and the effect of chronemics on relational communication. *Journal of Organizational Computing, 5*(4), 355–378. http://dx.doi.org/10.1080/10919399509540258.

Wolf, A. (2000). Emotional expression online: Gender differences in emoticon use. *CyberPsychology & Behavior, 3*(5), 827–833. http://dx.doi.org/10.1089/10949310050191809.