CMC CUES ENRICH LEAN ONLINE COMMUNICATION: THE CASE OF LETTER AND PUNCTUATION MARK REPETITIONS

Yoram M Kalman, The Open University of Israel, yoramka@openu.ac.il; Darren Gergle, Northwestern University, dgergle@northwestern.edu

Abstract

Decades of IS literature focus on the relationships between the use of lean forms of online communication, and various task and outcome variables. This paper suggests the need to examine the underlying assumption which classifies text-based media as inferior and lacking in social, relational and affective richness relative to traditional forms of communication. We review the evidence for this from the computer-mediated communication (CMC) literature, and point to the extensive evidence that lean media conveys these socio-emotional cues. We then focus on the lack of research on the mechanisms which facilitate this transmission of complex cues using only keyboard strokes. We review several descriptive studies which attempted to classify these cues, and suggest the term "CMC cues" to describe them. We suggest that in order to understand the mechanisms by which CMC cues operate, we must explore large, ecologically-valid and diverse samples of unobtrusively collected messages. We then focus on a single category of cues: character repetitions. These include repetitions of letters and repetitions of punctuation marks, cues which could emulate some of the characteristics of spoken nonverbal communication. These repetitions are studied in the Enron Corpus, a large collection (~500,000) of e-mail messages sent by and to employees of the Enron Corporation. Findings on the usage of this cue are discussed, and the findings are compared to findings on character repetitions in blogs and in Twitter micro-blogs. We conclude that letter repetitions often emulate spoken nonverbal cues, and that the frequency of character repetitions in CMC is highly variable and context dependent. A framework for further systematic research of CMC cues is presented.

Keywords: CMC, nonverbal, cues, media richness, lean CMC, repetitions.

1 INTRODUCTION

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 are not conclusive. Nevertheless, it is clear 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 (for a review see 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). This introduction reviews the evidence for the existence of these cues (termed *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.

1.1 The Cues We Use Online

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 (Doring & Poschl, 2008; Kalman & Rafaeli, in press; Sheldon, Thomas-Hunt, & Proell, 2006).

Another category of cues that has been researched 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, & 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 are labeled 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, the same authors, in collaboration with Postmes (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 few paragraphs 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.

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; 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 a subjective 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 methods did not allow for the measurement of distributions and the identification of regularities in the data. However, it is only through the identification of such patterns that we are able to elucidate the possible mechanisms that allow these cues to convey the socio-emotional information.

In this study, we employ novel 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 organizational e-mails. We explore the usage of letter repetitions and of punctuation mark repetitions. This cue (character repetitions) 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 were extensive enough to suggest the principles by which this cue operates in CMC. In this study we aim to collect a sufficiently large and diverse sample of character repetitions, and to suggest, based on these findings, the general principles by which these repetitions act as cues in CMC.

Before moving on to the research question, a few words on the definition of the term *CMC cues* as it is used in this paper to describe the cues which convey important social and relational information. 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 analogically to traditional nonverbal cues, to describe *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*.

1.2 The Research Question

In this study we examine the role of letter and punctuation mark repetitions (character repetitions) in e-mail messages. Like many other CMC cues besides chronemics and emoticons, character repetitions have not been systematically studied, and their usage is not well understood. The goal of the study is to understand the usage of character repetitions in an ecologically valid, large-scale sample.

The study was conducted in the context of a single dataset (the Enron Corpus, see below) which allowed maximal flexibility of search queries. The research question was:

How are letter and punctuation mark 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. An answer to the question *whether the letter repetitions are articulable or not* provides 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, part 2.1.2.

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 were 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 will provide some evidence as to the function that repetitions play and provide some insight into repetitions' relation to the more traditional communication functions of spoken language.

Further, we asked whether letter repetitions are a cue which is 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 *which of the users in our corpus use letter repetitions in their email messages*.

Lastly, we extended our analysis beyond letter repetitions that take form only within a word and explore the use of terminal punctuation marks which impact the interpretation of sentences. Punctuation marks are used according to grammatical rules, as well as to indicate qualities such as rhythm, direction, pitch, tone and flow (Truss, 2003, p. 70). The punctuation mark repetition is a cue which is similar to the letter repetition in some ways, but different in others. We examine *the frequency of repetitions of question marks, periods and exclamation points in email communication*.

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 post-processing as described in the following sections.

2.1.1 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, 2008). 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.

CorpusCruizer was used to identify the occurrence and location of repetitions of the 26 letters of the alphabet, as well as of exclamation points, question marks, and periods.

2.1.2 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 forwarded or a reply-to) 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 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 sound 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. russssshhhh, 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. 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.

A small number of messages were excluded from coding: (1) a joke 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 was repeated on average 3.6 times in the dataset (range: 1-54) due to replication of identical messages. 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.

Root word occurrences	Number of independent occurrences	Number of independent 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

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

Of the 815 entries, the vast majority (767) were classified as articulable. For example: “freeeezing”. Only 12 were classified as inarticulable (e.g. “buttttttt”), and 36 were classified as including both articulable and inarticulable repetitions (e.g. “lllooonnnngggg”). The most prevalent part of speech of the entries was interjection (376). Some examples include “Yipeeee”, “Aughhhhh” 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 which 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. 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 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.

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)

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

3.2 Frequency and usage of punctuation mark repetitions in the Enron Corpus

The number of occurrences in the Enron dataset of three punctuation marks and their repetitions are detailed in Table 3: exclamation points, question marks and periods.

Number of repetitions	Exclamation point (!)	Question mark (?)	Period (.)
1	203617	558193	7230048
2	11353	40950	11390
3	32890	15753	68310
4	2827	6278	17647
5	1257	22856	6290
6	555	1806	3092
7	266	4952	1763
8	208	971	1192
9	205	771	772
10	126	537	731
11+	800	11265	4659

Table 3. Frequency of single and repeated punctuation marks in the Enron Corpus.

4 DISCUSSION

This study explores the frequency and usage of character repetitions in e-mail communication. 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. A comparison to repetitions in blogs and in Twitter postings suggests that the frequency of usage of the character repetition is highly variable and context dependent. We note that the variability and context dependency of the repetitions are analogous to hallmark attributes of traditional nonverbal cues. The discussion is then extended to the insights this study sheds on findings reported in the SIP and SIDE literature. We conclude with suggestions for further research of CMC cues, to elucidate how CMC cues facilitate the conveyance of complex and often subtle social and interpersonal messages.

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 syllable in spoken conversation:

> “I was in an electronics store the other night... Panasonic has 9" Portable DVD player (like your sony) with an 8 hour battery... $999.00 US. It is sweeeeeeet”. (This is a direct quote from the Enron Corpus. In this paper 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.......vvvvvrrrrooooommmmm.............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 frieeeeeeeeennnnnd”

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:

> “uggggghhhh!!! 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 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 The frequency of character repetitions is highly variable and context dependent

Following a discussion of emoticons in e-mail, Naomi Baron makes the comment “Are additional paralinguistic cues really necessary for sending satisfactory email messages? Probably not” (Baron, 2000, p. 242). While we agree that these cues might not be *necessary* for sending satisfactory e-mail messages, our research reveals extensive use of letter and punctuation mark repetitions in email messages. How extensive is this usage, and what can that teach us about character repetitions? Overall, the frequency of character repetitions is in the same order of magnitude of previous descriptive studies of CMC cues (Blackman, 1990; Carey, 1980; Spitzer, 1986). Like other CMC cues, the frequency of repetitions ranges between rare and ubiquitous. For example, the frequency of letter repetitions in the Enron Corpus is low, while the frequency of punctuation mark repetitions is high. Although we have several thousand examples of letter repetitions from the Enron dataset, their frequency is rather low. We found almost 3,000 occurrences in more than 500,000 messages, about six repetitions in every thousand messages. The relative frequencies are also highly variable. While the relative frequency of the word *so* with repeated o’s in the Enron Corpus is low, the relative frequency of the repetition *hmmm* is very high, and the relative frequencies of repetitions in the root words *ah* and *oops* are only about one order of magnitude less than the frequency of the root word (Table 2).

In contrast with the low absolute frequency of letter repetitions in the Enron Corpus, the frequency of punctuation mark repetitions is quite high, though even within this category we find significant variance. According to Table 3, there were more than 50,000 appearances of a sequence of two exclamation marks or more, more than 40,000 of two question marks or more, and more than 115,000 of two periods or more (of which about half were an ellipsis). When we compare the ratio between the number of single appearances of a punctuation mark, and the number of occurrences of repetitions of that same punctuation mark (the relative frequency), we see that the highest ratio is in exclamation marks (one repetition for about every four appearances of a single exclamation mark), and the lowest ratio is in regards to repeated periods, which appear once or twice per 100 appearances of a single period. We conclude that punctuation mark repetitions are ubiquitous in the Enron Corpus.

4.3 Character repetitions beyond the Enron Corpus

So far we have discussed findings on repetitions in the Enron Corpus. This corpus is extensive, and CorpusCruizer allows a level of detailed analysis which traditional search tools do not afford. Nevertheless, it is important to study how findings from the e-mail based Enron Corpus relate to other media and audiences. An exploratory study of character repetitions in blogs and in Twitter micro-blogs reveals that character repetitions are not unique neither to the medium of e-mail, nor to the population of Enron employees around the turn of the century.

In a random sample of 10,000 Twitter postings ("tweets") that were collected at various times of the day during weekdays and weekends in February and March 2010, there were 1126 (11.3%) occurrences of letter repetitions of three or more consecutive identical letters. Since a single tweet might include more than one letter repetition, 200 of the above were randomly sampled and coded. 15 of them (7.5%) included letter repetitions. In the same sample, 13 (6.5%) included repeated exclamation points, 3 (1.5%) included repeated questions marks, and 38 (19%) included two or more consecutive periods. Since ellipses are sometimes inserted automatically into messages, this usage might not be fully under user control.

It is interesting to note the appearance of letter repetitions in Twitter messages. Given that users are limited in the number of characters they can use (140 characters per message), it could be expected that microbloggers will try to use the small amount of characters allocated to each message in as an economical a manner as possible. As already discussed by Herring & Zelenkauskaite (2009), the detection of extensive usage of repetitions despite strict length limitations points to the importance of repetitions to the users, and perhaps its value as a tool to communicate social meanings that are traditionally conveyed through speech (p. 26-27). Since letter repetitions do not add verbal information to the message, this abundance of letter repetitions in Twitter is in agreement with the proposal that these repetitions play an important role in communicating relational, social or affective messages.

Using Google Blog Search (http://blogsearch.google.com), the frequency of letter repetitions in the top five root words from the Enron Corpus was measured in blogs posted between 2000 and mid 2009. The analysis revealed millions of examples of bloggers using repetitions such as *hmm* or *hmmm*. For example, for every thousand occurrences of the word *so* in blogs, there were five or more occurrences of the same word with a repeated letter o. An even more complex picture emerges when we explore the relative frequencies of the repetitions in blogs longitudinally. The relative rate of appearance of some of the repetitions is relatively constant, while others seem to go in and out of fashion in the blogosphere. For example, in 2008, the root word *ah* appeared almost always with one h, and appeared with two to five h's in only 6% of the cases, while in 2004 the ratio was quite close to 1:1. This ratio increased since 2000, peaked in 2004, and has been dropping since. In contrast, the relative frequency of hmmmmm in relation to hmm has remained quite steady between 2002 and 2008.

In conclusion, we see that the frequency of repetitions is highly variable, and context dependent. The relative frequencies of letter repetitions range from those similar to that of the root word, to several orders of magnitude below it. These frequencies vary between different words and different media, as well as longitudinally.

4.4 Letter and character repetitions as CMC cues

Our findings on the presence and usage of letter and punctuation mark repetitions will not surprise most users of computer-mediated communication. They reveal that some repetitions are very common, some are less frequent, and some are very rare; and they show that letter repetitions often emulate paralinguistic cues used in spoken conversation. In rare cases there is no apparent equivalent in spoken conversation, in which case it is possible that the repetitions are used to provide visual emphasis. We also present evidence that when emotionally-laden interjections are used, repetitions are more likely to be employed. All of these findings are in line with our suggestion that character repetitions in CMC serve as cues that extend the lexical meaning of the words, add character to the sentences, and allow the fine-tuning and personalization of the message. These support the evidence already presented in the past by SIDE and SIP oriented studies Taken together, we can conclude that character repetitions are CMC cues.

4.5 Toward a theory of CMC cues

In the introduction to this paper we described studies which identified dozens of forms of CMC cues, as well as studies which demonstrate social and relational roles of specific CMC cues. Our findings on repetitions provide further support to the claim that CMC cues perform a role which is sometimes analogous to the role nonverbal cues perform in traditional communication, that of enhancing the socio-emotional richness of messages. Both CMC cues and traditional nonverbal cues are multi-modal, highly diverse, context dependent, ubiquitous, and are often inseparable from the verbal content of the message. They are used by communicators to encode and to decode social and relational messages. Given this deepening understanding of the role and function of CMC cues, it is surprising that theories which label text-based CMC as impoverished and with cues filtered out, are still employed in order to explain online communication. We wish to suggest that CMC cues are as important and central as nonverbal cues in traditional communication. In the next few paragraphs we highlight evidence for the tentativeness in the literature, and suggest that there has accumulated a critical mass of evidence to accept CMC cues as an inseparable and important component of CMC.

Lea and Spears (1992) talk in their first study about "particular [paralinguistic] cues that were introduced into the messages” (p.328), referring to spelling and typing errors, an exclamation mark, and ellipses. In their second study they say that "Paralanguage use was disliked by subjects", and "paralanguage users tended to be seen as more submissive, incompetent, cold and inhibited" (p.333), and in the general discussion they say that "for both [categories of users], paralinguistic cues (such as spelling/typing errors and punctuation marks) inserted into messages were associated with the formation of impressions of the personal attributes of the message senders" (p. 335). These quotes exemplify the perception that paralinguistic cues are an element that is added to the messages, i.e. that some messages include these cues, and some do not. This perception might stem from the methodologies employed in the two studies, but it is evidence for a narrow and limited conceptualization of CMC cues. Enough evidence has accumulated since these studies have been published to conclude that the cues are not inserted or introduced into messages. Rather, the cues are there all the time, and in every message. There is no such thing as a paralanguage user, since all users employ paralinguistic cues. A message that has no spelling mistakes and typos is not devoid of cues. The fact that there are no such mistakes is a cue by and of itself. A message with only commas and periods and no exclamation marks or ellipses does not lead to less impression formation, but rather to a different impression than alternative messages. In the spirit of the famous "you cannot not communicate" (Watzlawick, Beavin, & Jackson, 1967) quote, we can paraphrase and say "you cannot not provide cues", and by extension to say that every CMC message contains many CMC cues. The choice not to use a specific cue such as a letter repetition or a punctuation mark repetition in a blog or an e-mail message is as important as (and often indistinguishable from) the decision to use it.

Another example is from a later paper already discussed above: Postmes, Spears and Lea (2000). In this study, "messages were counted for the number of paralinguistic markers in the text. Specifically, counts were made of ellipses, inverted commas, quotation marks, and exclamation marks. Sequences of question marks, exclamation marks, periods, or other symbols were double weighted" (p. 349). Thus, the researchers treat a specific list of CMC cues as "paralinguistic markers" which are simply counted and then aggregated. Why should exclamation marks be counted, but question marks not (unless they are repeated in a sequence), and why should a single period or dash not be counted, but a sequence of the same punctuation marks counted? This methodological treatment of some cues as cues that count, other as cues that count more, and others as cues that do not count at all is more evidence for the early narrow conceptualization of CMC cues as a single and additive component of the text, which presence can be measured on a single linear scale. It is time to move beyond this conceptualization, and state that CMC cues are abundant, diverse, and an inseparable aspect of every CMC message. Like research on traditional nonverbal communication, CMC research must not ignore these cues, and like sophisticated traditional communication practitioners, CMC users should give significant weight to the cues they use in text-based online communication.

A good example of the negative consequences of ignoring the richness of cues in CMC has been recently scrutinized by Walther, DeAndrea, Tong, Kim and Lenore (2009) who replicated a study by Epley and Kruger (2005). The Epley and Kruger study concluded that email correspondence is more likely to perpetuate stereotypes than spoken conversation. The work by Walther et al. exposes the questionable implicit assumption in that study, which is that transcriptions of a spoken conversation are the equivalent of e-mail based exchanges. This implicit assumption is not valid since such a transcription eliminates the abundance of CMC cues which would serve a role equivalent to the cues of traditional communication. It is no wonder that a spoken conversation that has been stripped down to only its verbal content, sans CMC cues, will lead researchers to the erroneous conclusion that only spoken conversation can convey these subtle cues. Walther et al. show that following the correction of this assumption, the negative bias of e-mail is removed. Moreover, in accordance with the hyperpersonal communication framework, under these conditions CMC can even be superior to spoken communication in its ability to lead to positive impressions.

4.6 Further research

Given the ubiquity of CMC in our society, further study of CMC cues is as important as the study of nonverbal cues in traditional communication. What form should this study take? We suggest a methodological plan of research that starts with the descriptive, and leads to experimental studies which explore the hypotheses suggested by the descriptive studies. The descriptive studies should explore individual cues in large, ecologically valid samples, in an attempt to identify the parameters along which the cue can vary. Then, the relative frequencies of various forms of the cue should be measured, to identify both common and rare forms of the cue. Lastly, the descriptive studies should explore actual usage of the cue, and the co-occurrence of specific cue clusters. Experimental studies should then explore the way the cues are used for encoding and decoding messages, and the interaction between different cues, as well as between the lexical meaning of the message and CMC cues. Like the relationship between the verbal and nonverbal content of spoken messages, CMC cues too are likely to be inseparable from other (e.g. verbal) components of the message.

4.7 Limitations of the study

Despite the range of texts explored in this study, and the diversity it reveals, this preliminary foray into the realm of character repetitions in e-mail messages is still limited in several important ways. The first is that descriptive studies are inherently limited to inference from observed frequencies and usage. In order to establish the role of repetitions as CMC cues which have a role that is analogous to nonverbal cues in spoken conversation, we need to move beyond descriptions and interpretations that are based on the personal experience of the researchers, as well as beyond evidence such as the part-of- speech analysis, the articulability analysis and the onomatopoeic words analysis. The work already carried out by Postmes, Lea and Spears, suggests that character repetitions may have influential outcomes. However, that preliminary work treated these cues in the aggregate (i.e. along with many other cues that were changed) and as a unidimensional additive construct. Based on the findings in this study, such work needs further elaboration and study.

This work focused only on three types of punctuation marks, and did not look at repetitions of other punctuation marks (e.g. dashes) and of other non-letter characters (e.g. asterisks). Further analysis of the corpus will help in elucidating the role of these in the emails. Similarly, this study did not explore the location of the repeated letters in the word, or the location of words with repetitions in the sentence or in the paragraph. It would be interesting to explore whether these are distributed evenly. For example, it appears as if more of the inarticulable repetitions appear at the end of words, rather than at the beginning or the middle. It is possible that the dynamics of text production make it easier to repeat a terminal letter than one that is in the beginning or middle of a typed word.

4.8 Conclusion

Letter and punctuation mark repetitions, as well as other CMC cues, have been studied and described by researchers in the past. These pioneering studies resulted in several nomenclatures for the cues that enrich the socio-emotional content of CMC messages. Several experimental and quasi-experimental studies related to SIP and SIDE have shown the role of such cues in conveying social, affective and relational messages. Despite the important role of these cues, most of the research to date has focused on chronemics and on emoticons while ignoring a wide spectrum of other cues which are more difficult to measure and whose role in communication is apparently more challenging to elucidate. This study focused on character repetitions, and explored it in an unprecedentedly large, diverse, and ecologically valid sample of e-mail messages, in an attempt to further understand its prevalence and usage. Several thousand occurrences of letter repetitions and several tens of thousands of punctuation mark repetitions in about 500,000 email messages were analyzed. The frequencies and usage of the repetitions were explored, showing that letter repetitions were much more likely to be articulable, to appear in interjections, and to appear in onomatopoeic words. Our findings on the prevalence and usage of repetitions not only in the Enron Corpus but also in blog and Twitter postings, exemplify the ubiquity and importance of CMC cues: they allow text-based CMC to convey the subtleties and nuances that were, in the past, assumed to be only in the realm only of traditional communication and of high resolution audiovisual multimedia. We show the abundance of evidence for the existence of dozens of categories of CMC cues, and outline the steps for a systematic study of the abundance, usage and effects of these cues, a study which is analogous to the exploration of the role of nonverbal cues in traditional communication. Based on this analogy, it is expected that CMC cues will emerge as sophisticated and significant components of CMC messages.

5 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.

6 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.

Baron, N. S. (2000). *Alphabet to Email*. New York: Routledge.

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.* Ph.D., Florida State University.

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*. Paper presented at the Proceedings of the 18th annual meeting on Association for Computational Linguistics, Philadelphia, Pennsylvania.

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

Cohen, W. W. (2005). Enron Email Dataset Retrieved October 27, 2008, from http://www.cs.cmu.edu/~enron/

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

Daft, R. L., & Lengel, R. H. (1986). Organizational Information Requirements, Media Richness and Structural Design. *Management Science, 32*(5), 554-571.

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*, 256-274.

Derks, D., Bos, A. E. R., & Grumbkow, J. (2007). Emoticons and social interaction on the Internet: the importance of social context. *Computers in Human Behavior, 23*(1), 842-849.

Doring, N., & Poschl, S. (2008). Nonverbal cues in mobile phone text messages: The effects of chronemics and proxemics. *The Reconstruction of Space and Time: Mobile Communication Practices*, 109.

Epley, N., & Kruger, J. (2005). When what you type isn't what they read: The perseverance of stereotypes and expectancies over e-mail. *Journal of Experimental Social Psychology, 41*(4), 414-422.

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. doi: 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), 1.

Kalman, Y. M., & Rafaeli, S. (in press). Online pauses and silence: Chronemic expectancy violations in written computer-mediated communication. *Communication Research*.

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

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.

Lea, M., & Spears, R. (1992). Paralanguage and social perception in computer-mediated communication. *Journal of organizational computing, 2*(3&4), 321-341.

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.

Mines, M. A., Hanson, B. F., & Shoup, J. E. (1978). Frequency of occurrence of phonemes in conversational English. *Language and speech, 21*(3), 221.

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.

Postmes, T., Spears, R., & Lea, M. (2000). The formation of group norms in computer-mediated communication. *Human Communication Research, 26*(3), 341-371.

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

Python Community. (2008). Regular Expression Syntax Retrieved October 27, 2008, from http://www.python.org/doc/2.5.2/lib/re-syntax.html

Sadler, J. D. (1971). Onomatopoeia. *The Classical Journal, 67*(2), 174-177.

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*, 1385-1395.

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.

Spitzer, M. (1986). *Writing style in computer conferences.* Paper presented at the IEEE transactions on professional communication.

Sproull, L., & Kiesler, S. (1986). Reducing social context cues: electronic mail in organizational communication. *Management Science, 32*(11), 1492-1512.

The Oxford English Dictionary. (2009)*The Oxford English Dictionary. 2nd ed.*: Oxford University Press. Retrieved from http://oed.com/.

Truss, L. (2003). *Eats, Shoots & Leaves*. London: Profile Books.

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., & D'Addario, K. P. (2001). The Impacts of Emoticons on Message Interpretation in Computer-Mediated Communication. *Social Science Computer Review, 19*(3), 324-347. doi: 10.1177/089443930101900307

Walther, J. B., DeAndrea, D., Tong, S., Kim, J., & Lenore, E. (2009). *Computer-mediated communication versus vocal communication in the amelioration of stereotypes: A replication with three theoretical models*. Paper presented at the NCA 95th Annual Convention, Chicago, Il.

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*, 355-378.

Watzlawick, P., Beavin, J. H., & Jackson, D. D. (1967). *Pragmatics of human communication. A study of interactional patterns, pathologies and paradoxes*. New York: W.W. Norton & Company.

Wolf, A. (2000). Emotional expression online: Gender differences in emoticon use. *CyberPsychology & Behavior, 3*(5), 827-833.