

The Reading Matrix © **2009** Volume 9, Number 2, September 2009

Short Message Service (SMS) Texting Symbols: A Functional Analysis of 10,000 Cellular Phone Text Messages

Robert E. Beasley Franklin College

ABSTRACT

The purpose of this study was to investigate the use of symbolic expressions (e.g., "BTW," "LOL," "UR") in an SMS text messaging corpus consisting of over 10,000 text messages. More specifically, the purpose was to determine, not only how frequently these symbolic expressions are used, but how they are utilized in terms of the language functions that they signal. The results of this investigation suggest that text messaging symbols are most frequently used to identify people and their relationships to other things (including other people). They also suggest that text messaging symbols are frequently used to express amusement. In addition, the results suggest that the text messaging community has developed its own language culture in which closing expressions, long-form dialog, and correct spelling and grammar are viewed as inefficient and impractical. Finally, the results suggest that text messaging symbols are often used to signal descriptions of things that are important to the message sender.

INTRODUCTION

Text messaging has become a way of life for many in the 21st century. Indeed, people can be seen in malls, schools, just about everywhere using their cell phones to send character-based messages to their friends, classmates, family members, and co-workers. This form of communication has become especially popular among young people. For one, it permits them to communicate with others from just about anywhere. Secondly, it permits them to communicate silently, which can be beneficial in noisy environments, like bars, when having an effective conversation on a telephone would be difficult, or where extraneous communication must be done quietly, such as in a classroom. Thirdly, it permits them to communicate both synchronously (i.e., two-way communication is occurring simultaneously) and asynchronously (i.e., two-way communication is delayed), thus combining some of the benefits of telephone and e-mail communication. This technology has also given rise to a new language form in which abbreviated spellings, acronyms, and other shorthand notations are almost universally employed by its users. It is precisely these features and their use that are at the heart of this research. More specifically, the purpose of this study was to investigate, not only how frequently these symbolic expressions are utilized, but how they are utilized in terms of the language functions that they signal.

Communication theory asserts that a sender transmits a message to a receiver. This message contains one or more expressions, each of which signals a language function, such as "acknowledging," "expressing happiness," or "topic shifting." Natusch (2005) illustrates the notion of messages, expressions, signals, and language functions via the following exchange of messages.

Susan: What did you pay for your ticket to Seattle last year? (Inquiring) Matthew: Something like (estimating) \$900 round trip, I think. (Informing) Susan: That was a good deal. (Evaluating)

In this example, Matthew's first message contains both the expression "something like," which signals that the sender is "estimating" and the expression "\$900 round trip, I think," which signals that the sender is "informing." A taxonomy was developed by Natusch (2005) for analyzing such conversational exchanges. It consists of 76 language functions and is useful for analyzing simple exchanges like those above. This taxonomy, which was heavily utilized in this research, is summarized in Table 1.

Language Function				
Acknowledging	Expressing amusement	Offering		
Acquiescing	Expressing annoyance	Opining		
Adding	Expressing anxiety	Praising		
Admonishing	Expressing boredom	Probing		
Advising	Expressing contempt	Qualifying		
Allaying	Expressing desire	Recalling		
Apologizing	Expressing discomfort	Referring		
Approximating	Expressing dislike	Refusing		
Attributing	Expressing dissatisfaction	Reminding		
Challenging	Expressing embarrassment	Reporting		
Checking	Expressing exasperation	Requesting		
Clarifying	Expressing happiness	Restraining		
Closing	Expressing ignorance	Retorting		
Comparing	Expressing irony	Stalling		
Complying	Expressing shock	Stipulating		
Confiding	Expressing surprise	Suggesting		
Declaring	Expressing unconcern	Summarizing		
Defending	Greeting	Sympathizing		
Defining	Guessing	Teasing		
Denying	Identifying	Thanking		
Descriptor	Informing	Topic shifting		
Disagreeing	Inquiring	Urging		
Downplaying	Instructing	Warning		
Emphasizing	Inviting	Yielding		
Encouraging	Invocating			
Expressing affection	Justifying			

 Table 1. Natusch's Seventy-six Language Functions Taxonomy (Used with Permission)

RESEARCH METHODOLOGY

Symbol Table

For this research, a Symbol Table was utilized comprising over 600 commonly used text messaging, e-mail, and chat symbols as well as their associated expressions and the language functions that they signal. This table was adapted from Natusch (2005), who originally compiled the list from several websites, including <u>www.askoxford.com</u>, <u>www.techdictionary.com</u>, <u>www.devever.net</u>, and <u>www.wikipedia.com</u>. A number of universally used symbols (e.g., XO and ZZZ) as well as several commonly used symbols specific to the target corpus (e.g., U and UR) were also added to the table. Table 2 shows a small sampling of the data in the Symbol Table.

Symbol	Expression	Language Function
ATM	At The Moment	Descriptor
AWOL	Away Without Leave	Descriptor
BBR	Burnt Beyond Repair	Descriptor
BWQ	Buzz Word Quotient	Descriptor
DOTGOV	A government official	Descriptor
НА	Ha!	Expressing amusement
НАНА	Ha Ha!	Expressing amusement
MSG	Message	Identifying
U	You	Identifying
UR	Your	Attributing
XO	Kisses And Hugs	Expressing affection
ZZZ	Sleep	Descriptor

Table 2. Symbol Table Data Sampling

Corpus

The corpus used in this research consisted of 10,117 SMS text messages collected by the Department of Computer Science at the National University of Singapore (How & Lee, 2004). These messages were collected from three sources. The first source was a small group of 20 undergraduate students who together contributed 6,167 messages to the corpus. Having a small number of students contribute a large number of messages over an extended period of time permitted sufficient message depth per user. The second source was a larger group of 146 undergraduate students. Together, these students contributed 3,348 messages to the corpus, thus permitting sufficient message breadth and representing a diverse set of users. The majority of the students in these two groups were Singaporeans between the ages of 18 and 22. These students were made aware that their messages would be made public. To collect the messages for the corpus, the students were asked to upload up to 75 messages to a website for which they would receive nominal compensation. They were instructed to only submit conversational English messages that were either sent from or received by their cell phones. They were also asked not to upload repetitive messages. Finally, the third source was Yahoo's SMS chat website, which broadcasts live SMS chats of certain SMS chat rooms. From this website, an additional 602 messages were collected from an estimated 30 people. Repetitive messages and otherwise noisy data were filtered out by the collectors.

Ambiguity

In human language, ambiguity is unavoidable. By its very nature, language is a negotiation in which those who are interacting progressively question and clarify in an effort to reduce ambiguity. For this research, several types of potential ambiguity were identified. The first type, *length ambiguity*, refers to a symbol that is comprised of only one letter. Symbols exhibiting length ambiguity were omitted from the analysis because they may be used as initials (e.g., H for "Henry") in the target corpus. However, both U and Y were retained for the analysis because (1) they are uncommon initials and (2) they were used frequently in the target corpus to mean "You" and "Why?" respectively.

The second type, *symbol ambiguity*, refers to a symbol that either spells an actual word (e.g., AS and BAG) or spells an acronym that has another generally-accepted meaning (e.g., ATM and SOB). Symbols exhibiting symbol ambiguity were also omitted from the analysis due to the difficulty of determining the intent of the symbol.

The third type, *expression ambiguity*, refers to a symbol that has more than one possible interpretation (e.g., HAHA could mean "Laughter," it could mean "Having A Heart Attack," or it could have some other generally accepted meaning within the target corpus). When a symbol exhibited expression ambiguity, the most likely expression, based on a manual inspection of the corpus, was retained for the analysis. If likelihood could not be determined, none of the symbols were retained for the analysis.

The fourth type, *function ambiguity*, refers to a symbol that has more than one possible language function (e.g., GIAR, which means "Give It A Rest" could have a function of restraining, or it could have a function of expressing exasperation). When a symbol exhibited function ambiguity, the most likely function, based on a manual inspection of the corpus, was retained for the analysis. If likelihood could not be determined, none of the symbols were retained for the analysis.

The fifth type, *culture ambiguity*, refers to a symbol that has both a corpus-specific meaning and a more commonly accepted meaning within the larger text messaging culture (e.g., BOT means "Bought" within the target corpus, but its more commonly recognized meaning is "Back on Topic"). When a symbol exhibited culture ambiguity, the corpus-specific symbol was retained for the analysis.

Analytical Procedures

Four major steps were required to perform the analysis for this research. The first step was to remove all the non-alphanumeric characters (except the apostrophes, which were used in the corpus to construct contractions) from the SMS corpus. A Visual Basic program was written to accomplish this task. The second step was to parse the corpus, placing each word in its own row in the Word Table. A Visual Basic program was written to accomplish this task as well. The third step was to identify any potential ambiguities in the Symbol Table and resolve them appropriately (as described previously). The fourth step was to analyze the entries in the Word Table against the entries in the Symbol Table. This was accomplished through a series of SQL queries (i.e., joins).

It is important to note that the identification and resolution of potential ambiguities required an iterative process. That is, some symbols emerged as potentially ambiguous only after a given SQL analysis was performed. For each symbol found in the Word Table, the corpus was manually sampled (using the text editor's search function) to ensure that the symbol was being used as expected. If it was not, its ambiguity was resolved in the Symbol Table (if possible) and the corpus was reanalyzed. If the ambiguity could not be resolved, the symbol was not retained for the next analysis. Thus, steps three and four were repeated several times before the final results of the analysis were obtained. Figure 1 summarizes the analytical procedures required for this research.



Figure 1. Analytical Research Procedures

RESULTS

Descriptive Analyses

For this investigation, the frequencies and proportions of the language functions signaled by the symbols in the SMS corpus were summarized. Table 3 shows this summary.

Language Function	Frequency	Percentage
Identifying	4294	67.22%
Attributing	734	11.49%
Expressing amusement	625	9.78%
Approximating	200	3.13%
Probing	156	2.44%
Closing	151	2.36%
Descriptor	70	1.10%
Topic shifting	49	0.77%
Urging	25	0.39%
Thanking	20	0.31%
Acknowledging	18	0.28%
Referring	11	0.17%
Stalling	9	0.14%
Reporting	8	0.13%
Requesting	6	0.09%
Offering	3	0.05%
Allaying	2	0.03%
Invocating	2	0.03%
Defining	1	0.02%
Expressing affection	1	0.02%
Expressing surprise	1	0.02%
Praising	1	0.02%
Inquiring	1	0.02%
Totals	6388	100.00%

Table 3. SMS Corpus Summary of Language Functions, Signal Frequencies, and Proportions

In an effort to identify the specific symbols that signaled the language functions in the corpus, a simple drill-down technique was utilized. Table 4 shows the results of this drill-down.

Language Function	Symbol	Expression	Frequency	Percentage	
Identifying	U	You	4025	63.01%	
Identifying	MSG	Message	167	2.61%	
Identifying	PPL	People	54	0.85%	
Identifying	BF	Boy Friend	35	0.55%	
Identifying	GF	Girl Friend	11	0.17%	
Identifying	WTH	With	2	0.03%	
Attributing	UR	Your	734	11.49%	
Expressing amusement	HAHA	Ha Ha!	586	9.17%	
Expressing amusement	HA	Ha!	26	0.41%	
Expressing amusement	LOL	Laughing Out Loud	10	0.16%	
Expressing amusement	EG	Evil Grin	2	0.03%	
Expressing amusement	LMAO	Laughing My A** Off	1	0.02%	

Table 4. SMS Corpus Results of Drill-down Technique

Approximating	ABT	About	199	3.12%		
Approximating	FW	Few	1	0.02%		
Probing	Y	Why?	156	2.44%		
Closing	CYA	See Ya	118	1.85%		
Closing	BYE	Good Bye	29	0.45%		
Closing	CU	See You	0.03%			
Closing	TTFN	Ta Ta For Now	Ta Ta For Now 1			
Closing	TTYL	Talk To You Later	1	0.02%		
Descriptor	IC	Some kind of technology	17	0.27%		
Descriptor	CO	Conference or Company	10	0.16%		
Descriptor	SEC	Section	8	0.13%		
Descriptor	ZZZ	Sleep	7	0.11%		
Descriptor	WU	Some kind of sport	6	0.09%		
Descriptor	WT	Weight	5	0.08%		
Descriptor	JC	Junior College	5	0.08%		
Descriptor	SUP	Supervisor	3	0.05%		
Descriptor	SOL	Solution	2	0.03%		
Descriptor	SU	Some kind of food	0.03%			
Descriptor	OT	Overtime	Overtime 2			
Descriptor	ZZ	Sleep	Sleep 1			
Descriptor	ZZZZ	Sleep	Sleep 1			
Descriptor	DL	Download 1		0.02%		
Topic shifting	BTW	By The Way	By The Way 49			
Urging	ASAP	As Soon As Possible	25	0.39%		
Thanking	THX	Thanks	20	0.31%		
Acknowledging	OIC	Oh, I See	18	0.28%		
Referring	BAK	Back	10	0.16%		
Referring	BAC	Back	1	0.02%		
Stalling	LTR	Later	5	0.08%		
Stalling	L8R	Later	Later 4			
Reporting	BOT	Bought	8			
Requesting	PLZ	Please	Please 4			
Requesting	SOS	Help!	Help! 2			
Offering	FYI	For Your Information	For Your Information 3			
Allaying	NP	No Problem	No Problem 2			
Invocating	TC	Take Care	Take Care2			
Defining	AKA	Also Known As	1	0.02%		
Expressing affection	XOXO	Kisses And Hugs	And Hugs 1			
Expressing surprise	WTF	What The F***?	1	0.02%		
Praising	WOA	Work Of Art	1			
Inquiring	ASL	Age, sex, location	1	0.02%		
Totals			6388	100.00%		

Inferential Analyses

In addition to computing frequencies and proportions, two-tailed t-tests were conducted to determine whether or not significant differences existed between the proportions of any given pair of language functions. Table 5 shows the results of this analysis.

	Identifying	Attributing	Expressing Amusement	Approximating	Probing	Closing	Descriptor	Other
Identifying	-							
Attributing	11.303*	-						
Expressing Amusement	12.117*	0.516	-					
Approximating	16.574*	3.132*	2.635*	-				
Probing	17.233*	3.499*	3.010*	0.415	-			
Closing	17.302*	3.537*	3.049*	0.459	0.045	-		
Descriptor	18.640*	4.275*	3.808*	1.377	0.982	0.939	-	
Other	17.175*	3.467*	2.976*	0.377	0.038	0.082	1.018	-

 Table 5. Results of the Two-tailed, Pair-wise T-tests (*p<0.05)</th>

DISCUSSION AND CONCLUSION

The results of this research suggest that SMS text messaging symbols are most frequently used to *identify* people and their relationships to other things (including other people). Indeed, Table 5 indicates that text messagers use symbols associated with the identifying language function significantly more than the symbols associated with any other language function. Of the six symbols that signaled the identifying language function, five identified people—the message recipient (U), people in general (PPL), boyfriends (BF), girlfriends (GF), and combinations thereof (WTH). These five symbols were used 4,127 times, accounting for 64.61% of all symbol usage. The somewhat common use of the symbols BF and GF seems reasonable given the makeup of the subjects who generated the target corpus (i.e., mostly university students between the ages of 18-22). A manual inspection of the corpus revealed that text messaging was frequently used to arrange for personal contact (a relational activity) between message senders and receivers (usually to eat or exercise). In addition, the symbol UR was used to *attribute* something to the message recipient 734 times, accounting for 11.49% of all symbol usage. Like the symbols BF, GF, and WTH, UR has relational connotations. That is, it is directly relating something or someone to the receiver of the message (e.g., your homework, your friend).

The findings also suggest that text messaging symbols are frequently used by message senders to *express amusement*. The symbols HAHA, HA, LOL, EG, LMAO were used 625 times, accounting for 9.78% of all symbol usage. Interestingly, during his *manual* analysis of a 544 message SMS corpus, Thurlow (2003) developed the strong impression of an overriding

humorous and teasing tone. He asserts that humor is used to maintain an atmosphere of intimacy and perpetual social contact amongst text messagers.

In addition, the results of this study suggest that the text messaging community has developed, in a sense, its own language culture. This idea is consistent with reports in the popular media, which suggest that the text messaging culture is so pervasive that many students are using its abbreviated notation (sometimes referred to as IM-speak or Instant Messaging-speak) in their school essays, term papers, and other writing assignments (Associated Press, 2007). Four aspects of this culture are discussed next.

First, it appears that expressions that signal *closings* are not an essential requirement of the language culture of text messagers. Indeed, the symbols CYA, BYE, CU, TTFN, and TTYL were used only 151 times in the corpus (accounting for 2.36% of all symbol usage). Assuming that one closing is signaled at the end of each group of related messages (e.g., only one party says CYA), the average number of messages in a message group would be 10,117 messages / 151 signals = 67. Even if we were to assume more conservatively that both parties signaled a closing at the end of each group of related messages (perhaps our algorithms did not detect all closing signals), the average number of messages in a message group would be 10,117 messages /(151 * 2) signals = 33.5. Both of these numbers, of course, seem unreasonable. In fact, data collected from American undergraduate college students by the author suggests that the average number of text messages in a group of related messages is approximately 7.7 (n = 27; sd = 2.2). A visual inspection of the corpus also revealed that non-symbolic closings, such as "See you," "Take care," and even senders' initials were virtually non-existent. When closings were present, they typically took the form of symbols like BYE, CU, and so forth. The implication is that expressions that signal closings are significantly less common in text messaging communication than in other forms of discourse (e.g., face-to-face, telephone, letters, and e-mails).

Second, a manual inspection of the corpus revealed that the language of the text messaging community is highly abbreviated, and that this abbreviated notation extends far beyond the symbols identified for this research. In fact, many abbreviations seemed to be generated for convenience only—without concern for any kind of consistency that would benefit reuse. Fung (2005) recognized this phenomenon and developed an algorithm that can not only distinguish between abbreviations with omitted letters, abbreviations with phonetic substitutions, actual words, and acronyms but can translate the abbreviations into their long form equivalents with reasonable accuracy. Hopefully, Fung's research will lead to practical improvements in text messaging communication.

Third, the visual inspection revealed that there is little concern for correct spelling or standard grammatical form within the text messaging community. Indeed, the corpus was virtually void of complete sentences and proper punctuation. One likely contributing factor to this phenomenon is that "Singlish" is the first or second language of most of the young subjects that generated the target corpus. Singlish is an English-based hybrid language native to Singapore that consists of words originating from English, Cantonese, Malay, and other languages with hints of American and Australian slang (learned from television). Its syntax resembles that of southern varieties of Chinese (Wikipedia, 2007). Thus, some misspellings and non-standard grammar might be expected.

Fourth, the analysis revealed that text message senders sometimes adapted symbols with commonly accepted meanings to suit their own needs. For example, the symbol WT, which is commonly used to mean "without thinking" in the larger messaging community, was used to mean "weight" in the target corpus, and the symbol SUP, which is commonly used to mean

"what's up?" in the larger messaging community, was used to mean "supervisor" in the target corpus. One possible explanation for this finding is that the message senders were simply unaware of the more common uses of these particular symbols. The implication is that the use of symbols is somewhat specific to the culture (e.g., age group, interests) of the network of text messagers.

Evidently, to the text messager, closing expressions are inefficient, long-form dialog is uneconomical, and correct spelling and grammar are impractical. Thurlow (2003) notes that the notation used by text messagers is characterized by several non-standard linguistic forms, including shortenings (i.e., missing end letters), contractions (i.e., missing middle letters), clippings (i.e., dropping the final letter), acronyms (e.g., BTW meaning "by the way"), initialisms (e.g., U meaning "you"), letter/number homophones (e.g., B4 meaning "before"), phonological approximations (e.g., "woz up?" meaning "what's up"), non-conventional spellings (e.g., UR to mean "your"), and misspellings and typos. This phenomenon is likely due to the awkwardness and inefficiency of the standard ISO keypad found on most of the cell phones that host the SMS technology (where three or four characters share a single key).

This problem has not escaped the attention of researchers, of course. How and Kan (2005) experimented with the arrangement of the characters on the keypad in an attempt to reduce the inherent ambiguity of keys with multiple characters. In addition, they developed an improved predictive algorithm that reduces the number of keystrokes required for a given word to be automatically completed. When both techniques were tested together, they found text input efficiency improvements of approximately 22% over today's standard predictive techniques. Similarly, Gong and Tarasewich (2005) studied various character-to-key mapping configurations where the characters were assigned to differing *numbers* of keys. In particular, they studied the effects of various constrained keypad designs (i.e., alphabetical order was not maintained across the keys) versus their unconstrained counterparts (i.e., alphabetical order was not maintained across the keys). They found that constrained keypad designs were easier for novice text messagers to learn and use.

The results of this investigation also suggest that text messaging symbols are often used to signal *descriptions* of things that are important to the message sender. Of the fourteen symbols that signaled the descriptor language function, two were used to describe technology (IC, DL), three were used to describe work (CO, SUP, and OT), three were used to describe school (SEC, JC, and SOL), three were used to describe sleep (ZZ, ZZZ, and ZZZZ), two were used to describe fitness (WU, WT), and one was used to describe food (SU). Although these fourteen symbols were used only 70 times, accounting for only 1.10% of all symbol usage, it was not surprising to find these items of interest given the age group of the message senders and their current status as university students.

Finally, it is possible that a text message corpus generated by native English speakers would yield different results from those describe here, especially with regard to spelling and grammar. Thus, future research plans include the compilation and analysis of such a corpus as well as a comparison of those results with the results of this research. This may provide some insight into the universality of the current study's findings. Other research plans include a comparative analysis of friendly and hostile online forums to determine the frequency with which symbols are used to signal positive and negative language functions. Such an analysis may lead to an approach for discerning the general focus and tenor of a corpus algorithmically.

REFERENCES

- Associated Press. (2007). Students use IM-lingo in essays. *CNN.com*. Retrieved February 9, 2007 from <u>http://www.cnn.com/2007/tech/02/09/chat.lingo.ap/index.html</u> (link no longer available).
- Fung, L. M. (2005). *SMS short form identification and codec*. Unpublished master's thesis, National University of Singapore, Singapore.
- Gong, J., & Tarasewich, P. (2005). Alphabetically constrained keypad designs for text entry on mobile devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 211-220). New York: ACM Press.
- How, Y., & Lee, M. F. (2004). NUS SMS Corpus. Retrieved December 28, 2006, from http://www.comp.nus.edu.sg/~rpnlpir/downloads/corpora/smsCorpus.
- How, Y., & Kan, M. (2005). Optimizing predictive text entry for short message service on mobile phones. In M. J. Smith & G. Salvendy (Eds.), *Proceedings of human computer interfaces international*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Natusch, B. (2005). Corpus of CMC texting signals. Retrieved January 15, 2007, from <u>http://geocities.com/bnatusch/textsigs1</u>.
- Thurlow, C. (2003). Generation txt? The sociolinguistics of young people's text-messaging. *Discourse Analysis Online*, 1(1). Retrieved February, 9, 2007, from <u>http://extra.shu.ac.uk/daol/articles/v1/n1/a3/thurlow2002003-04.html</u>.
- Wikipedia. (2007). Singlish. *Wikipedia.com*. Retrieved March 9, 2007, from <u>http://en.wikipedia.org/wiki/Singlish</u>.

Acknowledgement: The author wishes to express thanks to Dr. Timothy Garner, Associate Dean and Director of Institutional Research, for his assistance with the statistical methods used in this study.

Dr. Robert E. Beasley is professor of Computing and Chair of the Department of Mathematics and Computing at Franklin College in Franklin, IN. His research interests lie in the areas of online multimedia/hypermedia learning environments, computational linguistics, telecommuting, and energy and environmental engineering information systems.

E-mail: <u>rbeasley@franklincollege.edu</u> Webpage: <u>www.franklincollege.edu/pwp/rbeasley</u>