JOURNAL OF COMPUTER-MEDIATED COMMUNICATION

Journal of Computer-Mediated Communication

An Analysis of Language in University Students' Text Messages*

Fiona Lyddy
Francesca Farina
James Hanney
Lynn Farrell
Niamh Kelly O'Neill

Department of Psychology, National University of Ireland Maynooth, County Kildare, Ireland

Concerns over effects of 'textisms' on literacy have been reinforced by research identifying processing costs associated with reading textisms. But to what extent do such studies reflect actual textism use? This study examined the textual characteristics of 936 text messages in English (13391 words). Message length, nonstandard spelling, sender and message characteristics and word frequency were analyzed. The data showed that 25% of word content used nonstandard spelling, the most frequently occurring category involving omission of capital letters. Types of nonstandard spelling varied only slightly depending on the purpose of the text message, while the overall proportion of nonstandard spelling did not differ significantly. Less than 0.2% of content was 'semantically unrecoverable.' Implications for experimental studies of textisms are discussed.

Key words: Text messaging, Mobile phones, Language use, Linguistic, Literacy.

doi:10.1111/jcc4.12045

Text messaging, short message service (SMS) or 'texting' continues to be a popular means of communication, among young people in particular. A report by Lenhart, Ling, Campbell, and Purcell (2010) highlighted the rapid increase in text messaging in the United States, where 72% of teenagers use text messaging, compared to 51% in 2006. In a British survey, 52% of young people aged 11-18 (n = 1000), along with 28% of adults aged 18-65 (n = 2000), named texting as the most important form of communication that they use to stay in touch with friends (Mobile Life Report, 2008); for the young people surveyed, texting ranked above instant messaging (17%), e-mail (12%), calls via mobile (9%) or landline (10%), and letters (0%). A British survey of 2117 adults shows the increasing popularity of texting from 2005 to 2010, with 62% of those aged 16-24 preferring texting over other means of communicating with friends (Ofcom, 2011). Ling's (2010) cross-sectional analysis suggests that texting follows a life-phase pattern, with older teens and those in their early 20s making the most use of the medium, with usage dropping off with age.

The authors would like to thank Maria Bakardjieva and two anonymous reviewers for helpful criticisms and suggestions regarding data presentation and analysis.

*Accepted by previous editor Maria Bakardjieva

Texting is a fast, cost-effective, personal and nonintrusive means of communicating (see Ling, 2005). It is near-synchronous, and associated with distinctive styles of conversation and writing features such as 'textisms' (Carrington, 2004; Rettie, 2009). Textisms are language variants such as abbreviations and nonstandard forms of words (Crystal, 2008a,b; De Jonge & Kemp, 2010; Plester & Wood, 2009), and include features such as letter and number homophones (c for 'see', 2 for 'to'), contractions (txt for 'text') and nonconventional spellings (nite for night; see Kemp & Bushnell 2011; Plester et al., 2009; Thurlow & Brown 2003).

The limited analyses of text language that are available suggest that most of the language is standard and that distinctive or nonstandard forms occur alongside standard ones (Crystal, 2008 a,b). As Shortis (2007) points out, text messaging has "de-regulated what counts as English spelling rather than altered spelling itself" (p.21). Carrington (2004) borrows the term 'squeeze-text' to describe the principal features of text language. Words may be shortened to the minimum syllable length, often by removing vowels. Articles and conjunctions may be omitted, and numbers or letters may be substituted for graphemic units e.g., gr8 for 'great,' 4 for 'for,' 2 for 'to,' c for 'see,' or sum1 for 'someone.' Common phrases may be represented by acronyms (e.g. LOL, 'laugh out loud'). Capital letters might be omitted or used for emphasis. End-message punctuation may be absent. Various other abbreviations and nonstandard forms have been noted (see Carrington, 2004; Crystal 2008a; Drouin & Davis, 2009).

Letter/number homophones (e.g. l8r for 'later', or w8 for 'wait'), contractions, and emoticons are less frequently recorded in analyses of naturalistic text messages than media representations of text language would suggest (Ling & Baron, 2007; Thurlow, 2006). The variety and complexity of emoticons has, in particular, been exaggerated, with the 'smiley' and 'frown face' (:-) or ☺ and :-(or :- :- ☹) being the main emoticons used and accounting only for a modest proportion of message content (e.g. see Ling & Baron, 2007; Thurlow & Brown, 2003). Similarly, the main typographic symbol used in texts is an 'x' to signal affection, a convention commonly found in informal writing (e.g. see Thurlow & Brown, 2003). Thurlow and Brown's (2003) data show a low frequency of emoticons (:-)), typographic symbols (xxx), and letter/number homophones (gr8/great) in comparison to nonconventional spellings (nite/night), accent stylizations (ello/hello), and onomatopoeic spellings (yay!, haha), forms that suggest the influence of speech on the medium.

Estimates of the prevalence of textisms within text messages vary. Accounts agree that the majority of text language is standard form, and the nonstandard forms used are generally 'semantically recoverable' (Thurlow & Brown, 2003); after all, the texter will want to ensure that they are understood (Crystal, 2008a). The data contrast with media portrayals of text messages as an indecipherable code (see Thurlow, 2006). Ling and Baron's (2007) sample of American university students' text messages (191 texts) contained less than 5% textisms. Thurlow and Brown's (2003) analysis of 544 texts collected in Wales produced a higher estimate, with textisms accounting for 19% of total message content (see also Thurlow & Poff, in press). Crystal (2008b) suggests that about 10% of total message content is accounted for by textisms. There may be further variation across languages (Bieswanger, 2007; Döring, 2002 cited in Bieswanger, 2007; Ling, 2005).

The prevalence of textisms within text messages has been exaggerated in the media, with some descriptions treating text messaging as if it were a 'foreign' language (e.g., see Crystal, 2008 a,b; Jones & Schieffelin, 2009; Thurlow, 2006). Consequently there is much concern over the impact of the use of such forms on young people's literacy, a concern that is without strong empirical support (e.g., Plester, Wood, & Bell, 2008; Plester, Wood, & Joshi, 2009). Textism use arguably demonstrates an appreciation of the sounds of language (Crystal, 2008, a,b; Jones & Schieffelin, 2009; Plester & Wood, 2009; Tagliamonte & Denis, 2008; Thurlow, 2006). Plester and Wood's (2009) study of preteens found no negative effects on literacy for young users. Some studies have reported a positive effect of texting on children's literacy skills (e.g. Plester et al., 2009), although phonological skills may mediate some

of that relationship (see Wood et al., 2011). Some studies have noted negative effects on literacy skills, however. Rosen, Chang, Erwin, Carrier and Cheever's (2010) study of young adults showed a negative association between self-reported textism use and formal writing, while there was a positive association with informal writing. However, self-reported textism use was quite low in this case, and may or may not reflect actual use of textisms.

Concerns over effects on literacy have recently been supported by experimental research demonstrating processing costs associated with textisms. Experimental studies examining the reading of text messages have tended to focus on the nonstandard forms, even though much of the language in text messages is standard. For example, Berger and Coch (2010) compared event-related potentials (ERPs) in response to semantic anomalies in texted and standard English in young adults and found that responses to sentences which contained textisms mirrored those seen in nonnative language processing. However, the text stimuli were 'translations' of standard English sentences used in ERP studies and contained sequences not generally found in text messages (e.g. 'c@' for 'cat'). The context in which text messaging language is encountered was not considered when selecting the text stimuli. Berger and Coch (2010) note that the sentences used, "while relatively typical in standard English, are more unlikely to be on topics typically discussed in everyday texting" (p.145).

Similarly, Perea, Acha and Carreiras (2009) found a reading cost for sentences containing textisms when comparing young adults' eye movements when reading texted and standard Spanish. They selected words based on frequency in an SMS dictionary, but the sentences used were unlikely to be encountered in real text messages (for example, sentences included "finish the soup at once" and "we'll go to the concert on my bike").

Kemp (2010), using a textism translation/generation task, found that messages using textisms were faster to write than those in standard English, but they took nearly twice as long to read, and were associated with more reading errors. In this case, participants read preprepared sentences or wrote to dictation; again, the texted sentences exaggerated features found in real text messages. For example, a standard English message had a texted counterpart with 70% textisms in a 23 'word' message ("i h8 2 ask, but dnt 4get 2 txt me an aQr8 time to pu my frendz. thnx. def c u 4 dnr!": "I hate to ask, but don't forget to text me an accurate time to pick up my friends. Thanks. Definitely see you for dinner!"). Participants had difficulty reading some of the texts. BN (for 'being'), aQr8 ('accurate') and ez for 'easy' proved particularly problematic; in the latter example, as 'z' is pronounced 'zed' not 'zee' by this Australian sample, 'ez' would not seem a valid shortening, as Kemp notes. Kemp acknowledges that texts were longer and contained more textisms than would occur in real messages, and that texts between friends would likely produce less confusion. The low rate of intrusions of textisms into the conventional condition was also noted.

Other studies reporting reading costs associated with textisms have also used tasks that may overrepresent textisms. A study by Kemp and Bushnell (2011) used both a reading and a writing task. In the writing task, children (average age 11.5 years), were asked to type two messages that were dictated to them. In the conventional condition, the children were instructed to make sure that "all words [were] spelled correctly and with proper punctuation." In the 'textese' condition, they were instructed to type as they "would normally text a friend" (p.21). Kemp and Bushnell found that the proportion of textisms used by children instructed to write messages conventionally or using textese was 3% and 35%, respectively, supporting the notion that textism use is contextually controlled. But in a parallel reading task, messages were entirely conventional or almost entirely written with textisms. For example, the conventional sentence "When will we see you tonight? Because someone left a message about your friend being sick. Are you sick too?" became "Wen wil we c u 2night? Cause some1 left a msg bout ur frend bein sik. R u sik 2?" in the textese version. The finding that both speed and accuracy were compromised in the textese condition must be interpreted in light of these differences in stimuli.

Furthermore, the sentences devised were quite long; the conventional versions consisted of, on average, 20 words, or 107 characters, presented in 3 sentences.

Such studies make an important contribution to our understanding of the processing of spelling variants, but they may not always reflect real-world aspects of text messaging, such as the context of the communication, the sentence types and subject matter, and the proportion of textisms relative to standard forms. It is important to consider the extent to which the sentences used in such experiments reflect the linguistic characteristics of actual text messages.

Estimates of textisms vary in experimental studies, along with samples and methods. Kemp (2010) found that 50% of the content in generated text-messages was written as textisms, with some words always written by participants as textisms (e.g., words such as 'are,' 'message,' 'tonight' never occurred as standard spelling). In a parallel reading task, Kemp noted that the stimuli contained more textisms than would be found in naturalistic text messages in English, with 70% of the total content consisting of textisms. In Kemp and Bushnell's (2011) writing task, 35% of total content consisted of textisms. De Jonge and Kemp (2010) had their participants translate printed sentences with the instruction to type "as they would if sending the message to a friend." This produced a low proportion of textisms at 14% for young adults and 15% for teenagers. Studies with children have produced estimates of 34% for text generation tasks, in which messages were composed by the children themselves based on hypothetical scenarios (Plester et al., 2009), and 58% for dictated messages, where children translated preprepared standard English sentences into textese (Plester et al., 2008). Studies in which text sentences are generated by the researchers for use as stimuli produce an even higher proportion of textisms, as noted above. By contrast, analyses of naturalistic text messages have produced lower estimates, Thurlow and Brown's (2003) estimate of 19% in British undergraduates being among the higher of these. Mobile phone handsets and predictive texting capabilities have changed considerably since Thurlow and Brown's data were collected and there is a need for a more up-to-date analysis of the language that is used in text messaging.

As Kemp (2010) noted, experimental studies are limited by the lack of data on text messaging language. In the present study, we aimed to provide such data by examining the textual characteristics of text messages supplied by a young adult English-speaking sample. Data were collected in Ireland, where text messaging is a frequently used form of communication, particularly by young people. In the first quarter of 2010, mobile phones users in Ireland sent over 3 billion text messages, averaging 192 messages per subscription per month (ComReg, 2010). The focus of the current analysis was on textisms, defined as abbreviations, lexical shortenings and nonstandard spellings (see Kemp, 2010). The level of analysis focuses on the individual text message, rather than on the sender. We examined: (i) message length in terms of sentences, words and characters per text message; (ii) prevalence and types of nonstandard spelling; (iii) textisms as a function of message length; (iv) sender and message characteristics affecting spelling choice; and (v) word frequency. The aim of the analysis was to establish, in a reasonably large naturalistic dataset, the proportion of textisms used, context effects on textism use and the amount of language that was, to use Thurlow and Brown's (2003) term, 'semantically unrecoverable' to the objective reader. Furthermore, the variety of textisms used was examined by means of a frequency analysis of the text messages.

Method

Data generation

Text messages were collected from a convenience sample of 139 undergraduate students (99 women and 40 men) attending university in Ireland. All participants were Irish and English was their first language.

We recruited college students as participants in order to mirror the age range and educational profile of the young adults typically participating in the kinds of experimental studies outlined above (e.g. Berger & Coch, 2010; Kemp, 2010; Perea et al., 2009). The average age of participants was 22 years (SD = 4.1); the average age of the people the participants sent texts to, as reported by the participants, was 25 years (SD = 9.4). Each participant was asked to provide up to ten text messages sent in the previous week. Participants were informed that the research study was concerned with "the language used in text messages." Participants transcribed their texts verbatim onto paper; they were instructed (verbally and in writing) to carefully reproduce the original message, paying attention to spacing between words, punctuation marks and capital letters. They also provided other details, such as their age and gender, the age and gender of the message recipient, and their relationship to the participant (e.g., friend, family, workmate etc.). Each participant also selected the purpose of the text from a number of categories (e.g. seek information, reply, make arrangement, etc.). In contrast to previous work in which the researchers have coded the texts' function (e.g. Thurlow & Brown, 2003), we had participants supply this label, so that it fitted with their intention. Participants were asked to choose messages that were genuinely representative of those they generally sent. To encourage selection of representative messages, participants were given a code to use to obscure any private information during transcription (such as a name). A total of 133 text messages used this code, which suggests that this method was successful in allowing participants to include messages that they might otherwise have chosen not to share. Any remaining identifying information (a name and phone number in just one message) was removed. Participants chose which messages they wished to divulge and they were assured of the confidentiality and anonymity of their responses.

The messages were subsequently transcribed into an electronic document. A number of 'automated' text messages (such as forwarded SMS advertisements and 'chain' texts) were removed from the set. Three messages written in languages other than English were also removed. This resulted in a set of 936 text messages, with a total of 13391 words (tokens) and 676 nonword units (e.g., symbols, emoticons, and multiple punctuation marks); 72% of the messages were sent by women, and 66% of those receiving the texts were women. The majority of content consisted of text communication between young adults, mainly among friends. The following analysis treats the individual message as the unit of analysis.

Coding

Coding of texts was based on the typology used by Thurlow and Brown (2003) to analyze naturalistic data, following that of Shortis (2001). The categories are similar to those of De Jonge and Kemp (2010), which was adapted from Plester et al. (2009). The categories of nonstandard spelling were: Accent stylization; Contractions; Emoticons and typographic symbols; G clippings; Initialisms; Letter/number homophones; Misspellings; Missed punctuation (excluding end-message punctuation); Missed capitalization; Other clippings; Onomatopoeic/ exclamatory expressions; Nonconventional/phonetic spellings; Semantically unrecoverable words; Shortenings. Examples of each category, along with definitions, are presented in Table 1. After initial coding of the 936 messages, a sample of messages (approximately 10% of the set) was independently coded by a second rater. Interrater agreement was 98% across the categories noted in Table 1.

Results

The following analysis considers message length, the prevalence and types of nonstandard spelling, the effect of message length on textism use, the effect of sender and message characteristics (including the purpose/ function of the text message) and word frequency.

Table 1 Types, descriptions and occurrences of nonstandard types of spelling in order of frequency

Type	Definition	Examples	Number of occurrences	% of total nonstandard spellings
Missed capitalization	A words is spelled without appropriate capital letter	john, i'd	728	22.09
Accent Stylization	A word is spelled as it is pronounced in casual speech	wantz, wanna, gona, cuz, dis, ds	615	18.66
Letter/number homophones	A letter or number used to take the place of a phoneme, syllable, or word of the same sound	2 (to), 4 (for), l8r, u, r (are), c (see), gr8, ru, 2ni (tonight), 2gether	429	13.02
Missed punctuation	Omitted periods, and spelling with missing apostrophe	dont, cant, wont, ill	360	10.92
Contractions	Omitting letters from the middle of words	Txt, wknd, dnt, plz, bday, gng	168	5.10
Phonetic/ nonconventional spellings	A spelling of a word from sound	fone, nite, luk, buks, cum	183	5.55
G Clippings	Omitting the final g in a word ending 'ing'	goin, talkin, comin	171	5.19
Other clippings	Omitting other final letters	tel, I'v, hav, wil, com	156	4.73
Onomatopoeic/ exclamatory	A nonword sound-based exclamation	Ha, arrrgh, woohoo, yay	156	4.73
Shortenings	Omitting the end of a word, losing more than one letter	Prob, bro, mon, tues	138	4.19
Misspellings	Misspelled words	dont't (don't), juut (just), remeber (remember), thought (taught)	126	3.82
Initialisms	A word or group of words represented by initial letters	tb = text back, gf = girlfriend, poa = plan of action, nntr = no need to reply	39	1.18
Semantically unrecoverable	Words apparently not correct in current context, or where texter's intended word is not clear		27	0.82
Total			3296	100

1 *Message length* Word count was calculated manually for each individual text message, as automated word counts may not count lexical items that are joined by a punctuation mark (e.g. 'ok.see you' would be counted as two words, not three). The word count included words and lexical substitutions for words; for example, the text 'i wish i ws der' ('I wish I was there') contained five 'words' or lexical items. Additional characters, such as emoticons and punctuation marks, were not included in the word count. Character count was automated and included spaces. The average length of messages was 14.3 words (SD = 12.0); this is consistent with Thurlow and Brown's (2003) 14 words per message estimated using an automated word count. On average, messages used 70 characters, with considerable variation in character count (SD = 59.4). Seven percent of messages exceeded the notional 160 character per message limit. Messages consisted of, on average, 2 sentences (SD = 1.46), with a median of one sentence.
2 *Prevalence and types of nonstandard spelling* Of the 936 texts, 158 had completely accurate spelling, that is 17% of the messages contained no textisms or nonstandard forms of any kind (including missed capitals or punctuation, etc). Of the 13391 words in the dataset, 10222 (76%) were judged to have entirely standard spelling; this count used stringent criteria, including only those words that were spelled, punctuated and capitalized correctly. The requirement for capitalization was, in particular, rather a strict criterion; capitalization is often abandoned in text messages as it can require several keystrokes on some handsets.

Categories of nonstandard spelling, with definitions, examples and number of occurrences are shown in Table 1. A small number of acronyms (17) appeared across the set; these were standard acronyms (e.g. CD, DVD, BBQ) rather than initialisms and are therefore not considered to be nonstandard spellings here. Nonstandard spellings were defined as spelling deviations of the following types (see Table 1 for definitions): letter/number homophones; onomatopoeic words; contractions; shortenings; g-clippings; other clippings; initialisms; nonconventional/ phonetic spellings; accent stylizations; misspellings; missed punctuation; missed capitalization; unrecoverable forms. Onomatopoeic words are not necessarily nonstandard, and are often accepted in other forms of writing; however, to maintain a strict set of criteria so as not to underestimate the number of nonstandard forms, we included them here.

There were 3296 instances of nonstandard spelling, accounting for 25% of the total word content. This is a marginal overestimation, as a number of words (less than 1% of the set) were coded more than once (e.g. an item like '2moz' uses both a homophone and an accent stylization, while 'id' includes a missed capital as well as a missed apostrophe). Missed capital letters accounted for the highest proportion of nonstandard spellings, at 22% of the total. Consistent with Thurlow and Brown's (2003) analysis of British text language, accent stylizations were also a frequent deviation from standard spelling, accounting for 19% of all nonstandard spellings. Letter/ number homophones were also frequently employed, with 429 occurrences, accounting for 13% of the total nonstandard spellings. There was considerable variety within this category; for example, the word 'tomorrow' appeared as 2morrow, 2morro, 2moro, 2mro, 2mo, 2m, 2moz. The categories of nonstandard spelling are detailed in Table 1.

There were 676 instances of emoticons and typographic symbols, including multiple punctuation marks, which were the most prevalent type within this category. The use of a single x (n = 162) or sequence of x's (n = 100) was also prevalent. Of the 936 text messages, 107 (11%) used emoticons, with 136 emoticons in total. A small variety of emoticons was encountered, with smiley and frown faces accounting for the majority of instances. Of the messages containing emoticons, 81% were sent by women.

Table 2 Bivariate correlations between the number of words used and spelling types

	1	2	3	4	5	6	7	8	9	10	11
1. Words	–										
2. Standard spellings	.95**	–									
3. Homophone	.30**	.15**	–								
4. Onomatopoeic	.29**	.24**	−.04	–							
5. Shortening	.25**	.19**	.07*	.03	–						
6. Contraction	.23**	.07*	.28**	.02	.06	–					
7. G Clippings	.39**	.28**	.18**	.17**	.15**	.17**	–				
8. Other Clippings	.33**	.20**	.15**	.07*	.14**	.25**	.31**	–			
9. Initialisms	.20**	.17**	.07*	.01	.03	.04	.10**	.10**	–		
10. Nonconventional/ phonetic spellings	.27**	.10**	.42**	−.00	.13**	.33**	.21**	.21**	.08*	–	
11. Accent stylizations	.38**	.18**	.31**	.11**	.14**	.45**	.32**	.35**	.04	.40**	–

*p < .05; **p < .01.

A large number of nonstandard spellings were due to incorrect (midmessage) punctuation (360, or 11% of the total nonstandard spellings), the majority of which consisted of omitted apostrophes (349). Of the messages in which an apostrophe might have been used (611 messages), 43% were correct. The others omitted an apostrophe, most commonly within the possessive (e.g. writing 'Johns place' instead of 'John's place'). Other common errors included words such as I'm and it's.

Ten per cent of the messages contained more textisms than standard spellings. Just 27 'words' (or 0.2% of the total content) were semantically unrecoverable to the researchers; in at least some cases, these items may well have been understood by the texters themselves, or might have been clear had the context been reinstated.

3 *Does message length affect textism use?* There was a strong positive correlation between the number of lexical items (words and word substitutes) in the text message and the number of standard spellings present ($r = .95$, $p < .01$), as well as with each of the nonstandard spelling types (see Table 2), suggesting that longer messages (i.e. more words) provide more opportunity for both standard and nonstandard spellings. There was no correlation overall between message length and the proportion of the text that used standard spelling. However, of the messages that contained more textisms than standard spellings (10% of the set), a majority (67%) were short messages, below the average length of 14.3 words.

4 *Sender and message characteristics affecting spelling* Overall, there were more messages in our dataset that had been sent by women (677 messages from 99 senders, accounting for 72% of the data) than by men (259 messages from 40 senders), which must be taken into account when considering gender differences. There was no difference in message length, with messages from men using an average of 13.9 ($SD = 11.6$) words per message, while those from women used 14.5 words ($SD = 12.1$). Messages sent by women contained a higher proportion of nonstandard spellings; the effect size here was small, however. On average, men's messages consisted of 78% standard spelling ($SD = 20.3$); for women, 74% of content ($SD = 21.4$) was standard, $t(934) = 2.4$, $p = .017$, $r = .01$.

Some gender differences were suggested within the categories of nonstandard spelling, with small effect sizes in each case. Women were more likely than men to use (or to contribute texts containing) emoticons and typographic characters, such as multiple punctuation, $t(934) = 3.14$, $p = .002$, $r = .1$,

Table 3 T test comparisons (with Bonferroni correction) of language in messages sent by men and by women

	sent by men M (SD)	sent by women M (SD)	GP DIFFS	Significance	Effect size (r)
Number of words	13.9 (11.6)	14.4 (12.1)	NS	$p > .5$	.09
Emoticons and typographical symbols*	0.49 (1.1)	0.81 (1.5)	S	$p < .01$	.12
Standard spellings	11.05 (9.5)	10.9 (9.7)	NS	$p > .8$	.01
Homophones	0.38 (0.9)	0.49 (1.0)	NS	$p > .1$	.05
Onomatopoeic	0.17 (0.6)	0.16 (0.5)	NS	$p > .3$	.03
Shortening*	0.09 (0.32)	0.17 (0.5)	S	$p < .01$	.07
Contraction	0.12 (0.5)	0.2 (0.6)	NS	$p < .05$	.07
G clippings	0.2 (0.5)	0.18 (0.5)	NS	$p > .6$	.02
Other clippings	0.12 (0.4)	0.18 (0.6)	NS	$p > .1$	.05
Initialisms*	0.01 (0.1)	0.05 (0.2)	S	$p < .01$	.09
Non-conventional spellings	0.17 (0.4)	0.2 (0.6)	NS	$p > .3$	.03
Accent stylizations	0.54 (1.1)	0.71 (1.2)	NS	$p > .05$	.06

*significant at $p < .01$.
GP DIFFS Group differences.

shortenings, $t(934) = 2.1$, $p = .033$, $r = .03$, and initialisms, $t(934) = 2.9$, $p = .003$, $r = .1$ (see Table 3). While there was no overall gender difference in the use of accent stylizations, this form was more likely to occur in same-sex friend dyads i.e. men texting men and women texting women, $F(1,930) = 5.64$, $p = .018$, $r = .2$. No interactions between the gender of sender and receiver were observed for the other categories of spelling. However, the underrepresentation of data from men must be considered here; it may be that men text less overall or the texts collected here may not be entirely representative.

Participants also noted the purpose or function of the text from nine categories (e.g. to seek information, reply, make an arrangement, express humour etc; see Figures 1 and 2). The purpose was not given for 3 of the 936 texts. The most frequent reason given for sending the text was to seek information or make a request, accounting for 27% of all the text messages. Making arrangements (16.5%) and replying to messages (17%) were also frequent, as was sharing information (13%). This sample of text messages therefore would seem to underrepresent those sent for the purposes of maintaining or supporting relationships; for example, the greetings category accounting only for 6% of the total. This suggests a certain level of censorship when participants chose which messages to contribute. It may be that texts which support relationships were more personal and participants may have been reluctant to share them.

A one-way analysis of variance was conducted to examine differences in message content across the nine categories of text (see Figures 1 and 2). This identified differences across a number of the variables, including message length, $F(8,934) = 24.24$, $p < .001$, number of emoticons and typographic symbols, $F(8,934) = 17.9$, $p < .001$, standard spellings, $F(8,932) = 21.45$, $p < .001$, onomatopoeic expressions, $F(8,934) = 16.54$, $p < .001$, shortenings, $F(8,934) = 2.31$, $p = .031$, g clippings, $F(8,934) = 9.69$, $p < .001$, other clippings, $F(8,932) = 4.29$, $p < .001$, initialisms, $F(8,933) = 3.7$, $p < .001$, and accent stylizations, $F(8,934) = 8.019$, $p < .001$. Most of these differences simply reflected variation in message length; post hoc tests showed that messages sent to 'share information' or that participants identified as having

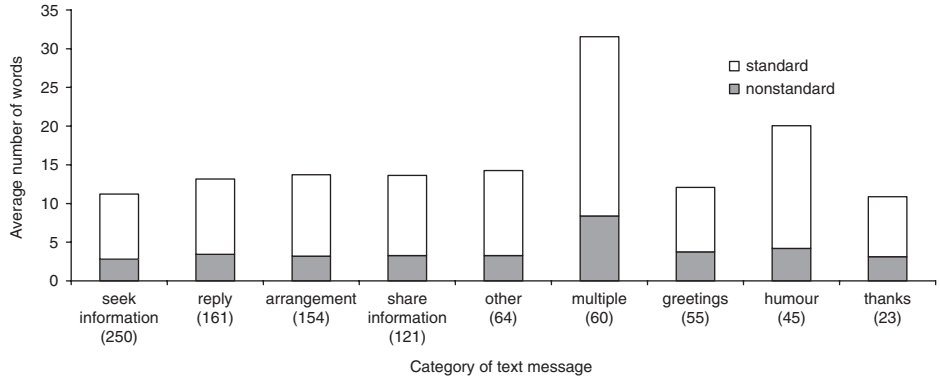

Figure 1 The average number of standard and nonstandard spellings for each category of text message (the number of texts in each category is given in brackets).

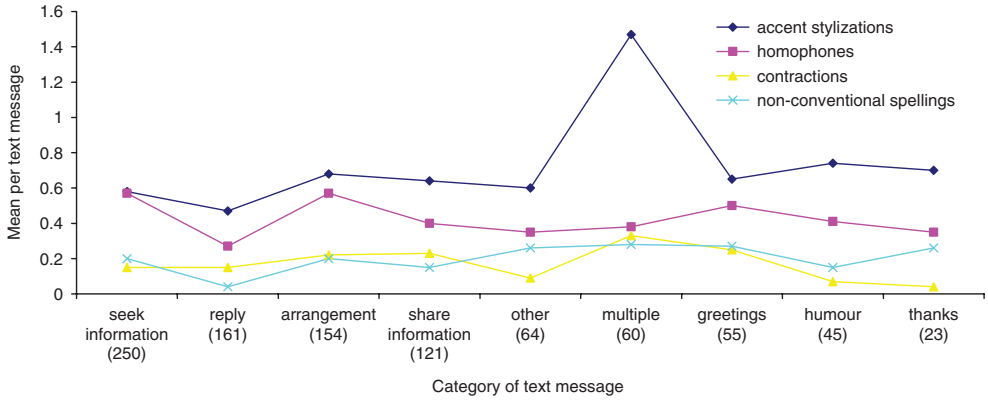

Figure 2 The mean number of the four most frequent types of textism across the nine categories of text message (the number of texts in each category is given in brackets). The 'multiple' functions category was reserved for use in cases where the participant could not identify one predominant function from those listed.

'multiple' functions were longer and contained more punctuation, emoticons, clippings, initialisms, and accent stylizations than other message categories (ps $< .01$).

There was no significant difference in the proportion of the text that used standard/nonstandard spelling across the categories, $F(8,933) = 0.75$, $p = .647$ (see Figure 1), despite differences in message length; the proportion of standard spelling varied from 71% in messages sent to thank someone to 79% in messages sent to share information. Similarly, shortenings, contractions, nonconventional spellings, and letter/number homophones did not differ significantly across the text message categories. These data might be taken to suggest that the texter's choice of a particular textism depends on the context of the message (see Figure 2); however, it may be that some texters used, or contributed to the study, particular categories more than others, which would affect the distribution of textisms across the categories.

5 *Word frequency* A computer program (TextSTAT; Hüning, 2002) was used to perform a simple frequency calculation (i.e. without tagging part of speech). The program identified 3073 distinct lexical items, 1896 of which received only one mention. The 100 most frequent words used accounted for 46% of the total content. Of the 100 most frequently used words, 16 were nonstandard spellings: u, i, hey, ya, yeah, im, 2 (for 'to'), d or D (for 'the'), 4 (for 'for'), b (for 'be'), Im, i'm, ur (for 'your'), goin, U. The numbers 2 and 4 appeared alone and within a number of words as homophones.

Given the size of our dataset (936 messages and 13391 words), comparison to established word frequency corpora is not likely to be reliable. Furthermore, text messaging tends to be used for specific purposes, and so the content would not be expected to be as general as in corpora of written or spoken language. However, comparison may still be useful in order to assess whether the most frequently used words in text messages are 'standard.' We compared the 100 most frequent words in our dataset with the 100 most frequently occurring words in the British National Corpus and the Brown (1984) British English corpus. The Brown (1984) corpus is based on 190000 words of spoken British English. The British National Corpus is a collection of 100 million words, mainly derived from written sources. These databases are not comparable; collections of written English tend to be far larger than those available for spoken English. However, given that we are only interested here in the most frequent words, it might still be informative to consider such sources.

Comparing our data to the high frequency items from the British National Corpus (see Kilgarriff, 2003), 58 words in the top 100 were shared, while comparing with Brown, 62 words of the most frequent 100 are shared. Fifty-five of the hundred most frequent words in our text messages appeared in both lists. The high-frequency words that appear in the text messages but are not found in the hundred most frequent words in the other corpora reflect the use of text messaging for social purposes, to make arrangements and engage with others; for example, words such as 'hi,' 'love,' 'meet,' and 'thanks' occur.

There are few detailed analyses of textism frequency in English available for comparison here. However, Kemp (2010) notes a number of high frequency textisms that were produced using a task in which participants were instructed to write down a dictated message using text abbreviations. Given that instruction, some words never occurred as conventional spellings; these words are shown in Table 4 below. Table 4 includes a consistency score for each word, as provided by Kemp (2010), which uses the weighted sums of the squares of the different textisms used for each word to give an overall consistency score as a percentage. Table 4 shows that, in the present analysis of naturalistic data, the proportion of standard spellings is higher and the consistency of textisms is lower than is reported by Kemp (2010) using a dictation task.

Discussion

This study set out to examine the textual characteristics of a large sample of text messages, with a view to informing experimental research using generated text messages. Consistent with previous research (e.g. Thurlow & Brown, 2003), this analysis found that 25% of the word content used some spelling variant, supporting the view that the majority of text messaging language is standard form (e.g., Crystal, 2008a,b; Thurlow & Brown, 2003). While our estimate is lower than the proportion used in the experimental studies highlighted above (which ranged from 35% to 50%), the proportion of nonstandard items, at 25%, is somewhat higher than other studies of naturalistic data have reported (e.g. Thurlow & Brown, 2003). The inclusion of missing capitals here has may have inflated our estimate; however, if we eliminate missing capitals, the proportion of nonstandard spelling is still high at 19%.

Missed capital letters were the most common nonstandard spelling, accounting for 22% of all such spellings. Consistent with Thurlow and Brown (2003) and De Jonge and Kemp (2010), accent

Table 4 Words which Kemp (2010) found were never spelled conventionally, using an experimental task, compared to the present naturalistic data. The consistency score uses the weighted sums of the squares of the different textisms used for each word to give an overall consistency score as a percentage. Percentages may not sum to 100 due to rounding

	Data from Kemp (2010)		Data from present study		
	% of textism consistency	Textisms produced (as % rounded to whole number)	% of conventional spelling	% of textism consistency	Textisms*produced (as % rounded to whole number)
are	100	r (100)	76.9	86	ar (7) r (93)
though	100	tho (100)	22.2	51	tho (57) dho (43)
through	100	thru (100)	100	–	
message	89	msg (94), mess (6)	71.4	24	msg (50) mess (50)
forever	64	4eva (79), 4ever (14) foreva (7)	100	–	
tonight	64	2nite (79), tnight (14), tnite (7)	67.5	24	tonite (20) 2ni (13) tnite (7) 2nyt (7) nite (33) 2nyte (7) 2 nite (7) 2n (7)
anyone	51	any1 (57), NE1 (43)	100	–	
back	49	bak (65), bk (24), bac (11)	82.8	52	bk (40) bak (60)
please	41	plz (59), plse (18), pls (18), pleas (5)	66.7	20	pls (25) plez (25) pleaseeeeee (25) plz (25)
thanks	28	thanx (35), thnx (35), thxs (12), tnx (12) thnks (6)	95.5	100	ta (100)
tomorrow	12	2morrow (14), 2moro (14), 2mro (14), tomoz (14), 2moz (10), 2morow (7), 2morw (7), 2mz (7), 2m (7)	53.4	23	tommorrow (4) tomorow (4) tomor (4) tomo (7) tmrw (4) 2morrow (4) 2morro (4) 2moro (43) 2mro (4) 2mo (4) 2m (11) 2moz (4) moro (4), 2 moro (1)

*Note: One misspelling (tommorrow) is included here. A number of items might be interpreted as a textism or as a misspelling (e.g., tnight, thnks, pleas).

stylizations also appeared frequently, accounting for 19% of nonstandard spellings. A sizeable proportion of nonstandard spellings consisted of some form of phonetic abbreviation, which may reflect a level of metalinguistic awareness, including phonological awareness, in texters (see also Plester, Wood & Joshi, 2009); however, as the textisms used are rarely novel, it may be that the originators of the forms, and not the end users who merely reproduce them, demonstrate these linguistic skills. A large number of nonstandard spellings were due to omitted mid-message punctuation (mainly apostrophes), which accounted for 11% of the total nonstandard spellings. Only 10% of the messages consisted of more textisms than standard spellings, in contrast to the stimuli typically employed in experimental manipulations comparing standard messages and those containing textisms (e.g. Berger & Coch, 2010; Perea et al., 2009). Our estimates of textism density also differ from studies in which the participant is instructed to generate, or translate, sentences using textisms (e.g., Kemp, 2010; Kemp & Bushnell, 2011; Plester et al., 2009).

Shortis (2007) notes that text language conventions have spread because the spelling used is ''linguistically coherent, logical, and creative in its orthographic principles and draws upon pre-existing conventions of nonstandard spelling'' (p.23; see also Thurlow & Brown, 2003; Crystal, 2008a). The present analysis would seem to support this assertion, with few 'semantically unrecoverable' items in the dataset (0.2% of the content). In at least some cases, these items may well have been understood by the texters themselves (which is why no examples are given here) or interpreted if the context were reinstated. The analysis of word frequency further supports the view that the language of text messages is largely standard, and makes use of existing conventions within informal writing. The 100 most frequent words used accounted for 46% of the total content. Of these words, 16 used nonstandard spelling and most, if not all, of these will be familiar from other forms of informal writing. The comparison to established corpora of written and spoken language showed substantial overlap with the text messages. Furthermore, comparison of the frequencies with those from an experimental study using generated text messages showed several differences.

However, the study has a number of limitations which need to be taken into account when considering these data. Participants contributed text messages of their choosing and their selection of messages may have been biased, particularly given that participants were aware that they were taking part in a study concerned with language in text messages. Our analysis of the function of the text message, as labeled by the participant themselves, contrasted with surveys on text messaging (e.g., Lenhart et al., 2010) and other naturalistic studies (Thurlow & Brown, 2003). In our dataset, messages used to make arrangements and get information were overrepresented compared to those sent to build relationships and support friendships. It may be that this type of text is more personal to the sender and participants may have been reluctant to contribute these to the study. The language we sampled will have been biased in this case.

The incidence of nonstandard forms in each type of text may also have been skewed, as some participants may have been willing to contribute messages of some types rather than truly representative text messages, over- or underinflating the proportion of textisms noted and perhaps accounting for the high frequency of emoticons here. Thank you messages seemed particularly formal, with the word 'thanks' spelled correctly and in full in the majority of cases. In addition, while our dataset consisted of a large number of text messages, and our analysis treated them as individual items, they were provided by a sample of 139 students, mainly women, and participants may have been motivated to take part in a study on text messaging based on their own texting practices.

Another potential bias is introduced by the transcription process. Participants transcribed their texts onto paper; these were subsequently transcribed into an electronic document. While care was taken at each stage, it is possible that errors could have been introduced through the transcription process, or that participants inadvertently 'corrected' their textisms during transcription.

A further limitation of the current study concerns predictive text settings. As we did not note whether the participants used predictive texting, our sample may not have been entirely representative; it may be that we sampled more, or fewer, users of predictive text than might be expected in the population. However, the focus here was on the language of individual text messages, as a reader might experience it, and in this context whether the sender chose to spell the word in full using predictive texting or by typing is less relevant. Furthermore, recording whether someone used predictive texting does not indicate how much of a text was written using the phone's predictive capabilities, a consideration that is further complicated by new smartphone applications which convert speech to (well formed) text for users.

The sample here consisted of university students. While this might be considered as a limitation in terms of generalizability, this sample was chosen in order to match the group typically participating in experiments comparing text messaging using textisms and standard spelling, as other studies have noted differences in textism use and effects depending on education (e.g., Rosen et al., 2010). The data are, however, likely to contain some features particular to this group, and to their locality.

Notwithstanding these limitations, this study highlights a number of issues that might inform experimental studies using laboratory-based analogues of real text messages. Text messages are used for particular purposes; reading of textisms is likely to be facilitated when these conditions are reinstated, providing a more accurate analogue of the real texting experience. The average length of the text message and the proportion of textisms might be considered in order to create a more ecologically valid comparison of standard sentences and those including textisms. Message length as well as the type of textism differed here depending on the purpose of the text message. The incidence of speech-based spelling variants also warrants consideration, as does the relatively small text vocabulary. Here, as in other studies (e.g., De Jonge & Kemp, 2010), missed capitalization was the most frequent form of nonstandard spelling. The high incidence of accent stylizations suggests that experimental studies need to consider local varieties of textism rather than selecting stimuli based on SMS dictionaries. When such factors are considered, it remains to be seen whether, or to what extent, processing costs affect reading or composition of text messages.

References

Berger, N.I., & Coch, D. (2010). Do u txt? Event-related potentials to semantic anomalies in standard and texted English. *Brain & Language*, 113, 135–148.

Bieswanger, M. (2007). 2 abbrevi8 or not 2 abbrevi8: A contrastive analysis of different space-and time-saving strategies in English and German text messages. *Texas Linguistics Forum*, 50. Retrieved October 28, 2010, from http://studentorgs.utexas.edu/salsa/proceedings/2006/Bieswanger.pdf

Brown, G.D.A. (1984). A frequency count of 190,000 words in the London-Lund Corpus of English Conversation. *Behavioural Research Methods Instrumentation and Computers*, 16, 502–532.

Carrington, V. (2004). Texts and literacies of the Shi Junrui. *British Journal of the Sociology of Education*, 25, 215–228.

Comreg (2010). Quarterly Key Data Report (June 2010). Retrieved on 21 November 2011 from http://www.comreg.ie/_fileupload/publications/ComReg1043.pdf

Crystal, D. (2008a). *Txting: The gr8 db8*. Oxford: Oxford University Press.

Crystal, D. (2008b). Txting: frNd or foe? *The Linguist, December* 2008, 8–11.

De Jonge, S., & Kemp, N. (2010). Text-messaging abbreviations and language skills in high school and university students. *Journal of Research in Reading*, doi: 10.1111/j.1467-9817.2010.01466.x

Drouin, M., & Davis, C. (2009). R U txting? Is the use of text speak hurting your literacy? *Journal of Literacy Research*, 41, 46–67.

Hüning, M. (2002). TextSTAT, version 2.8. Retrieved on 17 November 2011 from http://neon.niederlandistik.fu-berlin.de/en/textstat/

Jones, G., & Schieffelin, B. (2009). Talking text and talking back: ''My BFF Jill'' from boob tube to youtube. *Journal of Computer-Mediated Communication*, 14, 1050-1079.

Kemp, N. (2010). Texting versus txting: Reading and writing text messages, and links with other linguistics skills. *Writing Systems Research*, 2, 53–71.

Kemp, N. & Bushnell, C. (2011). Children's text messaging: Abbreviations, input methods and links with literacy. *Journal of Computer Assisted Learning*, 27, 18–27.

Kilgarriff, A. (2003). BNC database and word frequency lists. Retrieved November 17th, 2011, from http://www.kilgarriff.co.uk/bnc-readme.html

Ling, R. (2005). The sociolinguistics of SMS: An analysis of SMS use by a random sample of Norwegians. In R.S. Ling and P. E. Pedersen (eds.) *Mobile communications: Re-negotiation of the social sphere*, pp.335–350. London: Springer.

Ling R., & Baron N.S. (2007). Text messaging and IM: Linguistic comparison of American college data. *Journal of Language and Social Psychology*, 26: 291–298.

Ling, R. (2010). Texting as a life phase medium. *Journal of Computer-Mediated Communication*, 15, 277–292.

Mobile Life Report, 2008. Retrieved in 18th November 2011 from http://www.mobilelife2007.co.uk/Mobile_Life_2008.pdf

Ofcom (2011). UK adults' media literacy. Retrieved on November 21st 2011 from http://stakeholders.ofcom.org.uk/binaries/research/media-literacy/media-lit11/Adults.pdf

Perea, M., Acha, J., & Carreiras, M. (2009). Eye movements when reading text messaging. *The Quarterly Journal of Experimental Psychology*, 62, 1560–1567.

Plester, B., & Wood, C. (2009). Exploring relationships between traditional and new media literacies: Pre-teen texters at school. *Journal of Computer-Mediated Communication*, 14, 1108–1129.

Plester, B., Wood, C., & Bell, V. (2008). Txt Msg n school literacy: Does mobile phone use adversely affect children's literacy attainment? *Literacy*, 42, 137-144.

Plester, B., Wood, C., & Joshi, P. (2009). Exploring the relationship between children's knowledge of text message abbreviations and school literacy outcomes. *British Journal of Developmental Psychology*, 27, 145–161.

Rettie, R. (2009) SMS: Exploiting the interactional characteristics of near-synchrony. *Information, Communication & Society*, 12, 1131–1148.

Rosen L.D., Chang J., Erwin L., Carrier L.M., & Cheever N.A. (2010). The relationship between ''textisms'' and formal and informal writing among young adults. *Communication Research*, 37, 420–440.

Shortis, T. (2001). *The language of ICT: Information and communication technology*. London: Routledge.

Shortis, T. (2007). Gr8 Txtpectations: the creativity of text spelling, *English Drama Media*, June 2007, 21–26.

Tagliamonte, S. & Denis, D. (2008). Linguistic ruin? LOL! Instant messaging and teen language. *American Speech*, 83, 3–34.

Thurlow, C. (2006). From statistical panic to moral panic: The metadiscursive construction and popular exaggeration of new media language in the print media. *Journal of Computer Mediated Communication*, 11, 667–701.

Thurlow, C. & Brown, A. (2003). Generation Txt? The sociolinguistics of young people's text-messaging. *Discourse Analysis Online*, 1(1). Downloaded on 19 November 2010 from http://faculty.washington.edu/thurlow/papers/Thurlow(2003)-DAOL.pdf

Thurlow, C. & Poff, M. (in press). Text messaging. In S. C. Herring, D. Stein & T. Virtanen (eds), *Handbook of the pragmatics of CMC*. Berlin and New York: Mouton de Gruyter.

Wood, C., Jackson, E., Hart, L., Plester, B. & Wilde, L. (2011). The effect of text messaging on 9- and 10-year-old children's reading, spelling and phonological processing skills. *Journal of Computer Assisted Learning*, 27, 28–36.

About the Authors

Fiona Lyddy (fiona.lyddy@nuim.ie) is a senior lecturer with the Department of Psychology, National University of Ireland Maynooth. Her research interests focus on written language, including linguistic features of electronically mediated communication.

Francesca Farina (francesca.farina.2009@nuim.ie) is a doctoral student in the Department of Psychology at the National University of Ireland Maynooth. Her research focuses on spatial learning and memory.

James Hanney (hanney.james@gmail.com) is a Masters student in the Department of Psychology at the National University of Ireland Maynooth. His research focuses on health psychology.

Lynn Farrell (lynn.farrell.2010@nuim.ie) completed her undergraduate degree in psychology at the National University of Ireland, Maynooth. Her research interests include language and processing costs of text messages.

Niamh Kelly O'Neill (kellyonn@tcd.ie) completed her undergraduate degree in psychology at the National University of Ireland, Maynooth and a Masters degree in neuroscience at Trinity College Dublin. Her research interests include the psychology of language and the neuropsychological effects of brain injury.

Address: Department of Psychology, National University of Ireland Maynooth, County Kildare, Ireland.