

Cre8n Txt: A Rule-Based Approach to Textese

M Angel
University of Chicago

Abstract

This study investigates whether textism generation can be described through a set of rules – specifically, whether Textese has a context-sensitive, rule-based grammar that describes how textisms are created from their Standard American English counterparts. This was done by assessing 103 participants using a Textese translation task, a grammaticality acceptability judgment task, and a text-message submission task. Of 11 textism categories, rules were able to be created for eight of them. The remaining three categories had insufficient data on which to base substantial generalizations, although they were not inconsistent with a rule-based approach. Along with this, a singular predictable textism form was generated for 63% of English forms, while only 12.5% of English forms had more than one textism form, showing that textism generation is predictable considerably more often than it is not. With a majority of textism categories having generalizable rules and a relatively low rate of multiple textism forms for a single English form, it is likely that Textese has a rule-based grammar for generating textisms.

Introduction

Background

Text messaging has experienced a boom in popularity since the early 2000s, seeing a 62% increase in adult usage between 2005 and 2010, with even more growth since then (Lyddy et al., 2014). By 2014, 72% of teenagers reported that their preferred method of communication was texting (Lyddy et al., 2014). This rise to prominence of texting also came with the emergence of new, creative uses of language in text messages, referred to as Textese. Textese is the use of nonstandard orthography including, but not limited to, abbreviations, shortenings, and acronyms, most often used in informal computer-mediated communication, such as texting or instant messaging. A word written in Textese using nonstandard orthography is referred to as a “textism” (Shortis, 2007). An example of this is *lol*, an acronym of the phrase *laughing out loud*.

Academic studies of Textese often consider the sociolinguistics of texting (e.g., Thurlow & Brown, 2003) or examine how the use of Textese impacts literacy abilities (e.g., Coe & Oakhill, 2011; De Jonge & Kemp, 2012; Drouin & Driver, 2012). Other studies categorize textisms into “textism categories” (e.g., Crystal, 2008; Lyddy et al., 2014; Plester et al., 2009) based on similarities in their features and how they differ from Standard American English (SAE) spellings. Another subfield of research, in which this study is situated, is the investigation of how textisms are created (e.g., Kul, 2007; Kumar, 2012; McCulloch, 2015).

Previous Relevant Research

One of the earliest notable studies conducted on Textese is Thurlow and Brown (2003). The purpose of the study was to examine the sociolinguistics and discursive practices of young people texting. In discussing how textisms are created, the authors described them as “a creolizing blend of written and spoken discourse” (Thurlow & Brown, 2003). The authors give the example of the textism *novern*, an approximation of a regiolect pronunciation, specifically a Cockney accent, for the word *northern*. The study also found that this way of writing establishes an informal register in writing. This idea of texting – as the study put it, “write it as you say it” – was foundational for future research into how textisms are created.

A few years later, Shortis (2007) argued that textisms are vernacular forms of spelling, being rule-based, nonstandard orthographies. He claimed that Textese, or “Txt spelling,” was a significant shift toward conversationalism, or conducting communication using speech as if it were a casual conversation, and informalizing of writing, reflecting the findings of Thurlow and Brown (2003). Shortis (2007) explained that Textese draws upon pre-existing conventions of nonstandard spelling, being creative with the already existing principles of English orthography, such as swapping letters that are nearly homophonous (*teedge* for *teach*) or adding/removing letters to better approximate actual speech (*ope* for *hope*). Not only were textisms rule-based, but the rules that determined the Textese forms were already in use in English orthography. Tagg (2009) confirmed these findings and demonstrated that the patterns of respelling in Textese are consistent with existing patterns of spelling in English. These findings established the groundwork for the next decade of Textese research on the generation of textisms.

Subsequent studies set out to explain different categories of Textese. Kul (2007), expanding on the work of Shortis (2007), examined whether phonological principles could explain the deletion of letters in Textese. The study found that vowels are much more likely to be deleted than consonants, in line with the semiotic principle of “figure and ground,” or that “figures [vowels] tend to be foregrounded, grounds [consonants] tend to be further backgrounded” (Kul, 2007). Kul (2007) concluded that phonology did influence texting abbreviation.

Kumar (2012) examined contractions, or the removal of letters from the middle of a word, such as *txt* from the word *text*, to see if the removal of letters followed a pattern. In line with Kul (2007), the study found that vowels were much more likely to be deleted. Kumar (2012) concluded that there are two types of contraction: one in which all vowels (excluding initial sounds) are omitted, creating textisms such as *pls* from *please*, and one in which all letters except letters representing consonant sounds are omitted, creating textisms such as *msg* from *message*. Under the first type of contraction, *message* would be written as **mssg*. This subdivision of contraction showed that despite variation in the category, contraction generation could be explained through a rule-based approach.

Over the following decade, McCulloch (2015, 2020) investigated several areas of Textese. She concluded that shortenings, or the omission of the last part of a word, such as *prob* from the word

probably, were phonologically conditioned, and that shortenings are created by omitting everything past the stressed syllable in the word (McCulloch, 2015). Another finding was for the omission of capitalization in online contexts. A 2016 survey she conducted found that over half of the 500 participants (Twitter users) would manually undo autocapitalization to create the effect of missed capitalization (McCulloch, 2020). This shows a purposeful usage of Textese and that a rule is being applied quite literally, with an English form using a capitalized letter, and the author of the message going back and applying the rule of missed capitalization in order to achieve the textism form.

Even today, studies are still being conducted to explain Textese phenomena, such as Chen (2021), who investigated the syntactic placement of keysmashes, a sequence of seemingly random letters to indicate strong emotion, such as *AFGSAHF*. Yet with all this work being done, no study has looked at Textese as a whole, but rather has focused on individual parts of it. Previous findings of rule-based generation of textisms generally yield positive results, finding consistent patterns that can be generalized into rules. A holistic analysis of Textese would benefit linguistic studies of computer-mediated communication because such analysis could show whether Textese is fully rule-based. The alternative is that textisms exist as separate lexical items which substitute for conventionally spelled words. Previous scholarship tends to push back against the latter idea, as previous studies (e.g., Kul, 2007; Kumar, 2012) have taken the rule-based approach to textism generation. Therefore, a holistic analysis of textism generation in Textese is imperative to determine whether textisms are generated through rules or the lexicon.

Textism Category Categorization

The categorizations used to group textisms into textism categories vary between studies. The categorization used in this study is heavily inspired by Lyddy et al. (2014). This category set was chosen for the study because it provides descriptive definitions for each textism category, along with being inspired by previous literature such as De Jonge and Kemp (2010) and Plester et al. (2009). Table 1 defines all the textism categories used in this study and gives examples of each.

The key difference between this categorization and Lyddy et al.'s (2014) categorization is the merging of g-clipping (the removal of final g's) with other clippings (the removal of final letters that are not g). This was done to create a category that is broader in scope from these two nearly identical categories. The rest of the categorization from Lyddy et al. (2014) was chosen over De Jonge and Kemp (2010) and Plester et al. (2009) because it lends itself best to examining the generation of textisms from an English form. De Jonge and Kemp (2010) included extra categories such as symbols (i.e. :), @, *xoxo*) and a distinction between single letter and multi-letter homophones. Plester et al. (2009) also had a category for symbols. Symbols is a category that is not desirable for this analysis, as there is no English form attached to them.

Category	Definition	Examples
Missed capitalization	A word is spelled without an appropriate capital letter	<i>i'd, john</i>
Accent stylization	A word is spelled as it is pronounced in casual speech	<i>wantz, wanna, gona, cuz, dis, ds</i>
Letter/number homophone	A letter or number is used to take the place of a phoneme, syllable, or word of the same sound	<i>2 (to), 4 (for), l8r, u, r (are), c (see), gr8</i>
Missed punctuation	Omitted periods/punctuation marks	<i>I'm going to the store(.)</i>
Contractions	Omitting letters from the middle of words	<i>Txt, wknd, dnt, plz, bday, gng</i>
Phonetic spellings	A spelling of a word from sound	<i>fone, nite, luk, buks</i>
Clippings	Omitting the final letter in a word	<i>goin, yea, comin, tru</i>
Onomatopoeic	A nonword sound-based exclamation	<i>Ha, arrrgh, woohoo, yay</i>
Shortenings	Omitting the end of a word, losing more than one letter	<i>Prob, bro, mon, tues</i>
Misspellings	Misspelled words	<i>don't (don't), juut (just), remeber (remember)</i>
Initialisms	A word or group of words represented by initial letters	<i>tb = text back, gf = girlfriend, poa = plan of action</i>
Semantically unrecoverable	Words apparently not correct in current context, or where texter's intended word is not clear	<i>L.s.</i>

Table 1. Textism categories

The Present Study

The present study aims to expand upon previous research on rule-based generation of textisms. More specifically, the study looks to find if textism generation is influenced by linguistic factors that can be transcribed as a set of generalizable, context-sensitive rules that show the processes of how a textism is created, constituting a grammar of Textese.

For Textese to be considered fully rule-based, there must be a set of rules to explain the derivation of textisms from their English forms. Since these rules are context-sensitive – that is, they depend on the environment, or the surrounding letters/sounds, in which they apply – the structure of these rules will be modeled as a derivational phonological grammar (Halle, 1962):

$$A \rightarrow B / C _ _ D$$

Where A is the original grapheme(s) of the English form, B is the new grapheme(s) of the textism that A is rewritten as, and C and D are used to describe the environment before and after A in a word. If a rule is not context-sensitive, C and D will be omitted. Rules may also be subject to rule-ordering relationships, where multiple rules must be applied in a certain order for a specific textism to be generated.

Analysis was conducted by category, meaning each category was individually analyzed for patterns and generalizations that could be translated into rules. However, not all textisms fall nicely into a single category; rather, some fall into multiple categories. These will be called complex textisms and analyzed in terms of how the categories interact with each other in the form of rule-ordering relationships.

Methods

Data Collection

The data for the present study was collected in three parts: a translation task, a grammaticality acceptability judgment task, and text message submission. These tasks were designed to meet two goals. The first was to measure the variability of Textese, and the second was to identify how Textese is applied in both controlled and naturalistic settings.

The tasks were presented to participants via an online survey. The purpose of using an online survey was two-fold: Online surveys ensure that participants are at least familiar with internet technology and therefore are likely familiar with texting, and online surveys allow participants to type responses. Allowing participants to type their responses is vital for all three tasks, as typing may make participants more comfortable using Textese than if they were to write by hand. It also allows for participants to copy/paste their text messages for the text message submission task. Participants completed the survey either on their phones or on their computers. The difference between these two mediums is not of concern in this study, as each task asks for responses “as if you were sending a text message” for the translation and grammatical acceptability task, along with asking for copying/verbatim rewriting of sent text messages; thus, both mediums were allowed to complete the survey.

Participants

A total of 103 individuals participated in this study. Participants were chosen based on two criteria: age and current residence. All participants were between the ages of 18 and 22 years old. This demographic was chosen because Textese studies in the past have used similar age ranges (university students) for their research, with a mean age ranging from 19 to 22 years old (Kemp, 2010; Lyddy et al., 2014; Thurlow & Brown, 2003). The specific age range of 18 to 22 was chosen based on research that identifies this age range as when users of Textese use textisms most in their text messages (Ling, 2010).

All participants were also current residents of the United States. The scope of this study is restricted to American varieties of English, so restricting participation to those who live in the United States ensures that all participants are at least familiar with some American English. Most participants were from the Midwest, although a few were from the American South.

Translation Task

The first task participants saw was a translation task. The participants were given six sentences and were instructed to rewrite them “as if you were sending an informal text message.” Sentences were constructed using at least two English forms that are known to have corresponding textism forms found in previous literature (i.e., Crystal, 2008; Thurlow & Brown, 2003). The English forms vary across a wide array of Textism categories, including words or phrases with multiple, well-documented textisms, as well as English forms where a Textese rule could potentially apply. An example of the former would be /tu/¹ showing up in a word spelled as <to> and testing if it will be replaced with a <2>. This task yielded data on how Textese rules are applied in controlled phonological/orthographic environments. Since all participants received the same sentences, variability can be calculated by comparing participants’ responses. This task yielded 40 unique textism forms from 19 English forms. Comparing these textisms provided insight into whether there was a dominant textism form used for each sentence. Low variability and a dominant textism form provide evidence that textism generation is rule-based.

Grammaticality Acceptability Judgment Task

The second task given to participants was a grammaticality acceptability judgment task. This required the participants to study a list of textisms based on a certain English word or phrase to determine which of the textisms are grammatical and which are ungrammatical. Participants were shown each textism and asked to select it if they see it as a correct abbreviation of the base word or leave it unselected if it is an incorrect abbreviation. Grammars are constrained by abstract principles and rules (Mackey & Gass, 2005), meaning that there can be ungrammatical forms under a certain grammar if they do not follow the principles and rules of said grammar. This task aimed to determine if ungrammatical textisms are possible and, if so, what makes a textism ungrammatical. This can provide insight into whether Textese has constraints in its grammar and what would be considered a violation of the constraints. The presence of ungrammatical textisms would be evidence in favor of Textese being rule-based. If there are ungrammatical textisms, it implies that there are a set of rules that, when broken, yield an ungrammatical utterance.

The task consisted of six items: four words (*tonight, maybe, acknowledge, tomorrow*) and two phrases (*Be right back, Have a good day*). Three of the four words (*tonight, maybe, tomorrow*) and one of the phrases (*Be right back*) have attested textism forms, meaning that they are frequently used in Textese and have well-documented textisms. These words also have high variability (Crystal, 2008), meaning that there are multiple textisms for the words, so this task can also determine if all forms are grammatical, and if not, how much variability each item has in its grammatical textisms. Attested textisms were taken from Crystal (2008), Lyddy et al. (2014), and Thurlow and Brown (2003), or generated using documented Textism categories from Crystal

(2008), Kumar (2012), and Kul (2007). The remaining word (*acknowledge*) and phrase (*Have a good day*) do not have attested textisms. The items with no attested textisms, specifically *acknowledge*, were chosen for their relatively low frequency in SAE to test if Textese can apply only due to phonological/orthographic environments. Options for textisms were created for these items through application of patterns found in shortenings from Crystal (2008), application of rules for contraction described by Kumar (2012) and Kul (2007), and a common-sense application of a Textese rule, such as initialism by taking the first letter of every word. Each item in this task was also given a distractor textism as an option, which fits into none of the textism categories and therefore should be ungrammatical. If these distractor textisms, indicated in Table 2 with an asterisk, are marked ungrammatical by the participants, and if novel textisms are marked grammatical, it would be evidence in favor of Textese being rule-based, as it would show that there are productive rules which can be applied to novel words and phrases which yield a grammatical result. Table 2 shows all the tested words/phrases, along with the options for textisms associated with them.

Base Word/Phrase	Textism Options
<i>Tonight</i>	<i>2night, toni, 2nite, ton, tnght, *tght, tn</i>
<i>Maybe</i>	<i>mayb, mab, mAB, *myb, mb, mbe</i>
<i>Acknowledge</i>	<i>acknwldg, aknwlg, *aklg, ack, ackn</i>
<i>Tomorrow</i>	<i>tm, tmrrw, tmrw, 2morrow, *tw, 2mrw, tom</i>
<i>Be right back</i>	<i>brb, b right back, b rite back, b rht bck, *b rit b</i>
<i>Have a good day</i>	<i>hagd, hgd, have a gud day, hv a gd dy, *hae a go da, have a good dA, have a good da</i>

Table 2. Textism options for grammaticality judgments

The task was split into two sections: grammaticality and usage. The grammaticality section tests if the participant thinks the textism is grammatical (or “correct,” as it is put in lay terms for the survey). This provides information as to whether the participant would recognize and understand the textism and see it as an acceptable usage of Textese. The usage section tests to see if the participant would use or has used the specified textism in a text message. By eliciting which textism form(s) they have personally used, judgments were made as to which textism form is preferable for usage, and inferences were made as to why that textism is more used than other possibilities, despite both or all being deemed grammatical. It also showed whether any novel textisms would be sent in a text message, which would indicate that they are highly accepted despite just being created by the application of Textism categories.

Text Message Submission

The third and final task of the survey was text message submission. Participants were asked to retype or copy/paste verbatim between five and 10 text messages that they had recently sent that include nonstandard spellings. The requirement of including nonstandard spellings was to help elicit textisms without the need to define a textism for the participants. The purpose of this task

was to see how textisms are used in a naturalistic setting. A total of 663 text messages were received.

The text messages were analyzed by documenting all textisms present, along with providing a gloss of the SAE spelling equivalent and categorizing the textism into one of the 11 textism categories. A total of 1,669 textisms were documented from 712 English forms. While the previous two tasks aimed to see if it is possible for textism generation to be rule-based, the third task was used to derive said rules. The data was used to derive rules for each individual textism category, as each one affects English differently (Lyddy et al., 2014). These rules depend on contexts at the graphemic, phonological, morphological, and/or syntactic level.

Metrics Used

In order to quantify the distribution of textisms used for a single English form and to gain insight into whether optimal textism forms exist for a given English form, two metrics are used. The first metric used is Optimal Form Likelihood (OFL), which is calculated using the formula shown in Figure 1.

$$OFL = \frac{Mo - \mu_{(S-Mo)}}{N}$$

Figure 1. Formula for optimal form likelihood

Mo is the mode of the dataset, or the number of times the most frequent textism is used for a given English form. $\mu_{(S-Mo)}$ is the average of the dataset excluding the mode, or the average number of times a textism other than the most frequent textism is used. N is the total number of textisms written from the English form. This yields a number between 0 and 1 which describes how often the most used textism is written in comparison to other textisms. This is useful, as an OFL of 1 would show that a single textism is used every time for a given English form and is the optimal textism form. An OFL of 0 would show that there is no distinct optimal form, as each textism form is used an equal amount.

The second metric is usage. Usage is calculated by subtracting the number of times the English form appears as its SAE spelling from the number of times a textism is used for it. The result is either a positive or negative number. If usage is positive, it shows that textisms are used more often. If usage is negative, it shows that textisms are less commonly used than the corresponding English form. High usage shows that a given context is likely to create a textism.

These two metrics were calculated for each English form using a Python script. Utilizing these two metrics together provides insights into whether the English form has an optimal textism form, that is, a singular textism that is agreed upon. This can be expanded to looking at a large collection of English forms to see how many have optimal textism forms and how many do not. Table 3 shows the interaction between the two metrics and what it means for a given context to have a high or low OFL and usage.

	High OFL (≥ 0.5)	Low OFL (< 0.5)
High Usage (> 0)	Optimal textism form exists for English form	> 1 candidate form accepted
Low Usage (≤ 0)	Optimal form exists, but it is rarely used	No optimal textism form/Optimal form is English form

Table 3. OFL and usage interpretation

Results

Translation Results

In the six sentences used for the translation task, there were 19 English forms for which Textese was used. These yielded 40 unique textisms across the 19 English forms. An example of an English form that yielded more than one unique textism is *best friends*, producing the textism forms *bffs*, *besties*, and *bfs*. Table 4 shows the number of English forms that have a high/low OFL and usage.

	High OFL (≥ 0.5)	Low OFL (< 0.5)
High Usage (> 0)	5 (26.3%)	4 (21.1%)
Low Usage (≤ 0)	0	10 (52.6%)

Table 4. OFL and usage of translation English forms

Most of the English forms from the translation task (52.6%) do not have an optimal form, meaning that there is no single agreed-upon textism form for them but rather multiple textisms that are less preferred than the original English form.

Grammaticality Acceptability Judgment Results

Table 5 shows the OFL and usage of the six test items from the grammaticality acceptability judgments.

	Acceptability		Actual Use	
	High OFL (≥ 0.5)	Low OFL (< 0.5)	High OFL (≥ 0.5)	Low OFL (< 0.5)
High Usage (> 0)	1 (16.6%)	3 (50%)	1 (16.6%)	2 (33.3%)
Low Usage (≤ 0)	0	2 (33.3%)	0	3 (50%)

Table 5. OFL and usage of grammaticality judgments

Much like the translation results, there is only a single English form which has a single optimal textism form, or a textism form that is used a majority of the time (*be right back* having the optimal form *brb*). The remaining five English forms have either more than one accepted candidate form (multiple textism forms with high usage) or do not have a commonly accepted candidate form (no textism forms/textism forms with low usage). For the acceptability portion of the grammaticality judgment, the two test items that did not have any accepted candidate forms (low OFL and usage) were the two English forms without attested textism forms: *acknowledge* and *have a good day*. This a potential indicator that either Textese cannot be applied to novel contexts, or that the two contexts did not fit any of the rules of Textese, and therefore all candidate textism forms were considered ungrammatical.

The distractor textism for each test item had an average acceptability of 0.6472% (SD = 0.7236) and an average acceptability for actual use of 0.3236% (SD = 0.7236). This miniscule acceptance of the distractors shows that the participants were able to discriminate between acceptable and unacceptable textisms. It is also an indication that there is such thing as an ungrammatical textism.

One interesting finding is that the usage metric is higher for the acceptability portion than the actual use portion, or the section asking if texters would send a message using the textism given. Five out of the six items show this drop in the usage metric, with the item *maybe* changing from positive to negative between acceptability and actual use. This is an indication that there is likely to be more than one optimal textism form for some English forms, causing texters to have to choose a single textism from the optimal forms. The distinction between optimal forms could also be a contextual choice based on the message.

Text Message Submission Results

A total of 663 text messages were submitted by the participants for analysis. In these messages, 1,669 textisms across 712 unique English forms were used. Figure 2 shows the distribution of these textisms across the different textism categories.

From Figure 2, it is clear that missed punctuation is the most used textism category. The other most used textism categories include initialism, missed capitalization, and letter/number homophone. The OFL and usage results across the 712 English forms are shown in Table 6.

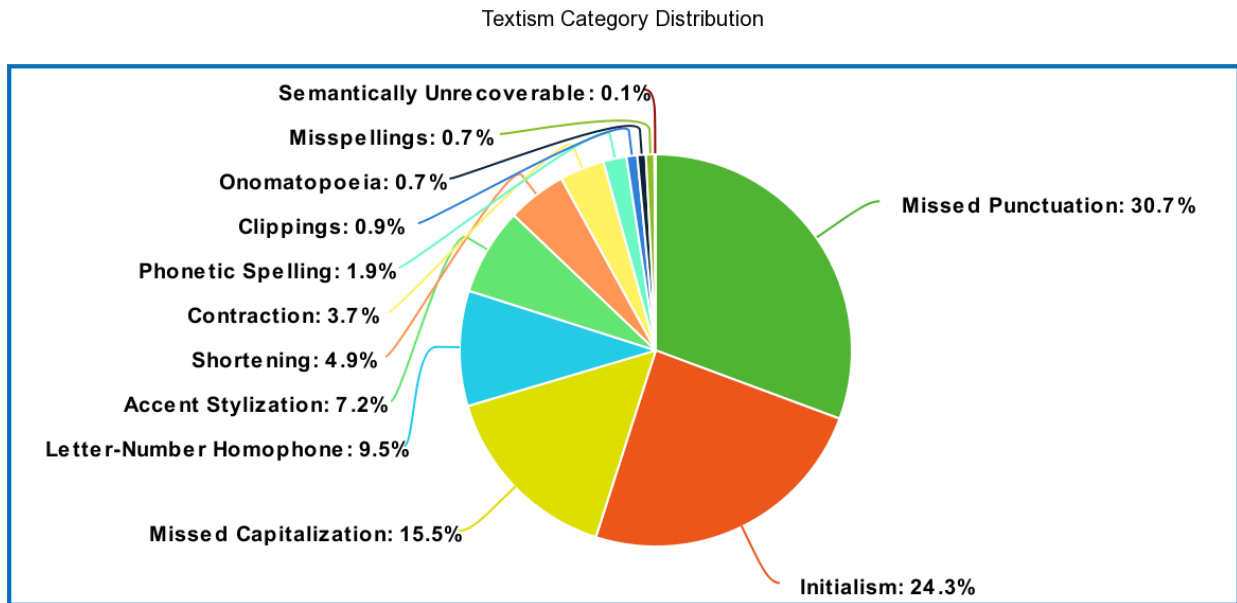


Figure 2. Distribution of textism categories

	High OFL (≥ 0.5)	Low OFL (< 0.5)
High Usage (> 0)	527 (74%)	27 (3.8%)
Low Usage (≤ 0)	63 (8.84%)	95 (13.36%)

Table 6. OFL and usage of text message submission

These results starkly contrast with those of the translation task and grammaticality judgments. Of the English forms, 74% have an optimal textism form with high usage, while only 13.36% of English forms lack an optimal textism form.

One factor that may be skewing the data is singleton English forms, or English forms that only have one instance of a textism associated with it. Since there is only one use of a textism for a given English form, its OFL may default to 1. An example of this would be the English form *Thinking*, which had only one instance of a participant using a textism, *thinking*. For comparison, Figure 3 shows the English form *you*, a non-singleton English form.

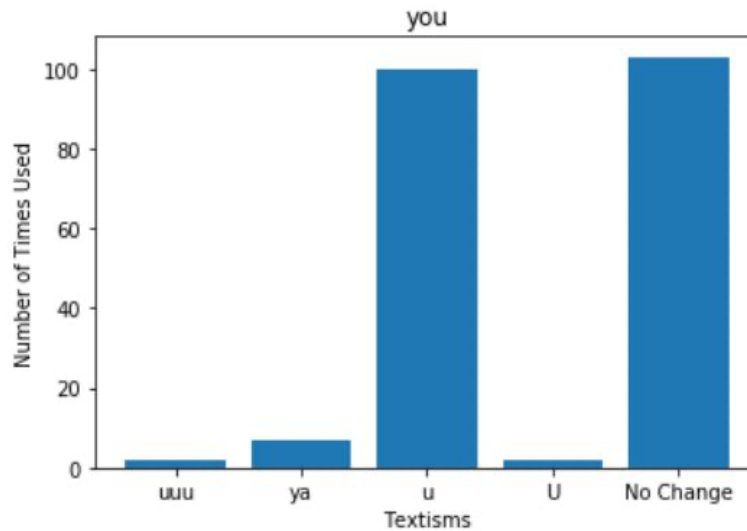


Figure 3. Textism distribution for the English form “you”

Since all the data gathered for this section is naturalistic, it was hard to control what data was received. This resulted in a large spread of textisms and many singletons. Of the 712 English forms, 496 of them are singletons. Because of this, it is necessary to run a separate analysis with the singletons removed to arrive at a fair estimate of how many contexts have optimal forms. Table 7 shows the OFL and usage of the 216 non-singleton contexts from this section. Figure 4 shows these English forms plotted by OFL and usage.

	High OFL (≥ 0.5)	Low OFL (< 0.5)
High Usage (> 0)	136 (63%)	27 (12.5%)
Low Usage (≤ 0)	9 (4.2%)	44 (20.3%)

Table 7. OFL and usage of non-singleton text message submissions

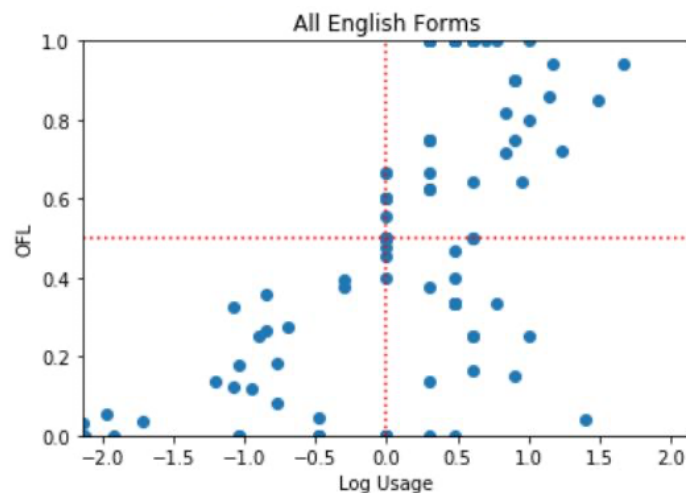


Figure 4. OFL and usage of all non-singleton English forms from text message submissions

Even with the singletons removed, 63% of English forms have an optimal textism form with high usage, while only 20.3% of contexts lack an optimal form. While these numbers are not as strong as the singleton numbers, they still show a considerable increase compared to the results for the previous two sections. This is promising evidence that Textese may be constrained by a rule-based grammar.

The following sections examine each textism category individually, analyzing the OFL and usage for both singleton and non-singleton contexts. Qualitative analysis was done to evaluate patterns across the textisms in each category to determine if they could be translated and generalized into rules.

Missed Capitalization

Missed capitalization made up 15.5% of the total textisms used. Missed capitalization, like missed punctuation, is unique among textism categories in that its context is just a single grapheme instead of a whole word or phrase. This also means that there is only one output possible; in this case it is decapitalization of the grapheme. Because of this, OFL defaults to 1.0 since there is only one possible form per English form. As these are not singleton contexts, this does mean that the form used is the optimal form. For missed capitalization and missed punctuation, only usage can change. For missed capitalization, there was a usage of -20. Since this is below zero, it is considered low usage, in that capitalization is used more often than missed capitalization. One possible explanation for this low usage is the use of autocorrect, which uses auto-capitalization. Kent and Johnson (2012) found that without auto-capitalization on smartphones, missed capitalization was used almost four times as often as correct capitalization. To use missed capitalization with autocorrect, one must go back and manually reverse what autocorrect did. The present findings of a somewhat low usage of missed capitalization (only 20 instances more usage of correct capitalization than missed capitalization) are close to the findings of McCulloch's (2020) survey, where slightly over half of respondents reported manually undoing autocapitalization.

Missed capitalization can be explained in only one rule. If a word starts with a capital grapheme, it is decapitalized. Figure 5 shows the rule written out formally.

$$\langle G \rangle_{[+cap]} \rightarrow \langle G \rangle_{[-cap]} / \# \underline{\quad}$$

Figure 5. Rule for missed capitalization

$\langle G \rangle$ is any grapheme (italics distinguish this $\langle G \rangle$ from the grapheme $\langle G \rangle$), $[+cap]$ means that the grapheme is capitalized, and $[-cap]$ means the grapheme is not capitalized. $\#$ indicates a word boundary, so in this case the rule only applies to graphemes at the beginning of the word. This rule applies to all English forms for missed capitalization.

While creating a rule to describe something as simple as decapitalization may seem trivial, it is important to do so for this study. As with McCulloch's (2020) survey, it shows intentionality in

using Textese, and that there exists a process for translating English forms into their respective textism forms.

Accent Stylization

Accent stylization made up 7.2% of all the textisms, with 122 instances of use. Accent stylization is when a word is respelled to mimic casual speech. It is also often used as a phonological approximation of regiolects (Tateman, 2015; Thurlow & Brown, 2003). Table 8 shows the OFL and usage of singleton and non-singleton English forms for accent stylization. Figure 6 shows these English forms plotted by OFL and usage.

	All Contexts (45)		Non-Singleton Contexts (25)	
	High OFL (≥ 0.5)	Low OFL (< 0.5)	High OFL (≥ 0.5)	Low OFL (< 0.5)
High Usage (> 0)	16 (35.5%)	4 (8.9%)	8 (32%)	4 (16%)
Low Usage (≤ 0)	7 (15.6%)	18 (40%)	1 (4%)	12 (48%)

Table 8. OFL and usage for singleton and non-singleton English forms of accent stylization

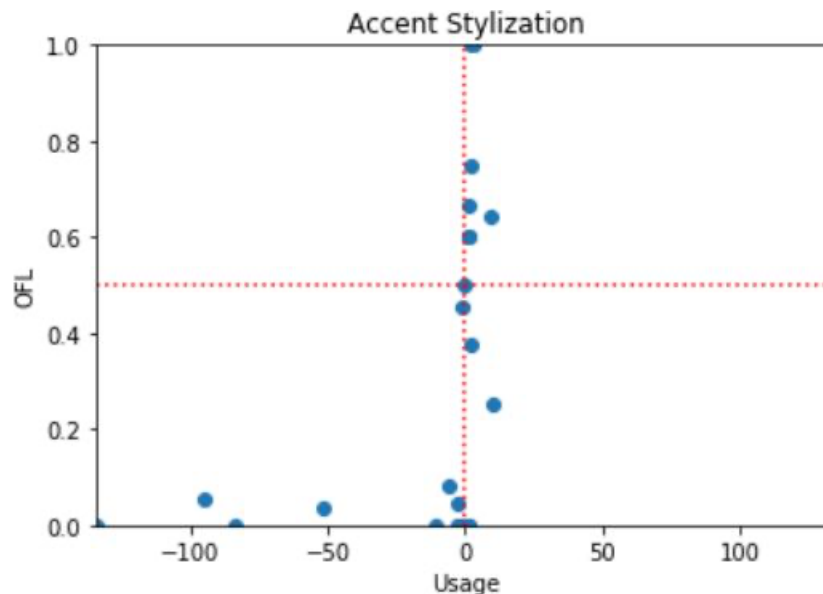


Figure 6. OFL and usage of non-singleton accent stylization English forms

As seen from Table 8, only 35.5% of English forms (32% of non-singleton English forms) have an optimal form, while 40% of English forms (48% of non-singleton English forms) have no accepted candidate forms at all. One possible factor that could explain this is variation in speech. Accent stylization is meant to mimic casual conversation or a regiolect, so texters from different geographic locations will use accent stylization differently since they have distinct regiolects. These regiolectal variations can mimic a certain dialect with sociophonetic factors encoded in the

orthography (Tateman, 2015), or they can be as specific as mimicking a particular person's way of speech (Tateman, 2016). These sociophonetic factors lead to high variation within accent stylization.

Despite this, two rules can be derived from accent stylization. The first, shown in Figure 7, describes word elongation through the repetition of graphemes to mimic lengthened speech.

$$\langle G_1 \rangle_{/p/} \rightarrow \langle G_1 * n \rangle_{/p/, n=3-6} / (\langle G_0 \rangle_{/p/}) \text{---}$$

Figure 7. First rule for accent stylization

In the first rule, $\langle G_1 \rangle_{/p/}$ represents any grapheme that corresponds to some phoneme p , $\langle G_1 * n \rangle_{/p/, n=3-6}$ represents the grapheme corresponding to phoneme p being reduplicated anywhere between three and six times (the data found the lower bound to be three and the upper bound to be six), and $\langle G_0 \rangle_{/p/}$ represents a grapheme preceding $\langle G_1 \rangle$ that corresponds to the same phoneme, as in a digraph or a grapheme that does not correspond to another phoneme, such as the English form *oh* turning into *ohhhhh*. In this example, the $\langle h \rangle$ is not recognized as $/h/$, but rather is an extension of the whole word $[ou]$. Another example of this would be *dream* becoming *dreaaaaam*, with the digraph $\langle ea \rangle$ corresponding to the phoneme $/i/$, but only the grapheme $\langle a \rangle$ is repeated. The final grapheme of a digraph/trigraph is the one that is repeatedly reduplicated. This indicates the prosodic feature of lengthening words in speech. $\langle G_0 \rangle$ is in parentheses, since it is not required if $\langle G_1 \rangle$ is the only grapheme that corresponds to its respective phoneme, as in it is not a digraph/trigraph. An example of an English form where $\langle G_0 \rangle$ would not apply is *lmao* turning into *lmaooooo*, since only the $\langle o \rangle$ grapheme corresponds to the phoneme being lengthened. The choice of the number of times to reduplicate is set at the lower bound of three to differentiate from words that naturally have double letters (i.e., *lot* becoming *loot*); a lower bound of three prevents unnecessary homography. The upper bound of six was found from the data in this study, though it may be possible to exceed this limit for ironic effect.

The second rule derives *wanna* contractions (i.e., *gonna*, *wanna*) and is shown in Figure 8.

$$\left. \begin{array}{l} \langle VC_1 C_2 \rangle \# \langle to \rangle \\ \langle VC_1 \rangle \# \langle to \rangle \end{array} \right\} \rightarrow \langle VC_1 C_1 a \rangle$$

Figure 8. Second rule of accent stylization

The left-hand side of the rule shows that it is initially two words (with the second word being $\langle to \rangle$, as in $\langle got to \rangle$ or $\langle want to \rangle$). The two rows on the left describe two instances of application, one where there are two consonants past the vowel (such as *want to*) and one where there is one consonant after the vowel (such as *got to*). These words are then merged into one, denoted by the removal of the space between them and the reduplication of the first consonant of the coda. This

rule is only used when the *wanna* contraction is followed by a verb, indicating that there are syntactic constraints on textisms as well. While Textese is not the first to use these types of contractions, they are a feature of Textese, and therefore cannot be left out.

Letter/Number Homophone

Letter/number homophones made up 9.5% of the total textisms, with 160 instances of use. A letter/number homophone is when a word or part of a word is replaced by a single homophonous symbol, such as using <2> for the word <too>, since they are both pronounced /tu/. Table 9 shows the OFL and usage of all English forms and non-singleton English forms. Figure 9 shows these forms plotted by OFL and usage.

	All Contexts (15)		Non-Singleton Contexts (8)	
	High OFL (≥ 0.5)	Low OFL (< 0.5)	High OFL (≥ 0.5)	Low OFL (< 0.5)
High Usage (> 0)	3 (20%)	0	0	0
Low Usage (≤ 0)	2 (13.3%)	10 (66.7%)	2 (25%)	6 (75%)

Table 9. OFL and usage for singleton and non-singleton English forms of letter/number homophone

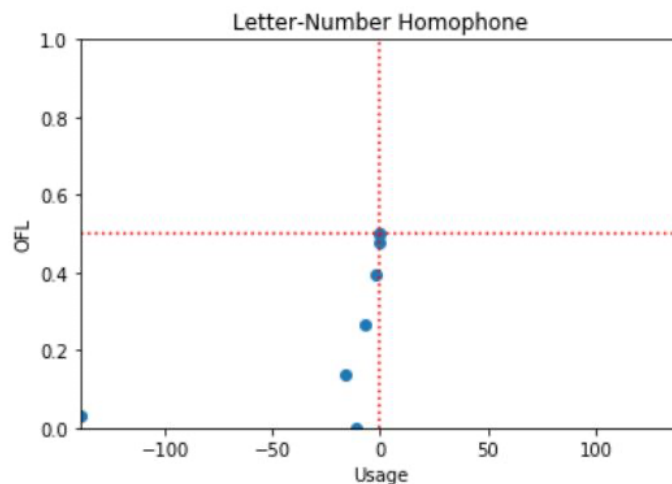


Figure 9. OFL and usage for non-singleton letter-number homophone English forms

One very noticeable characteristic of the letter/number homophone category is how little it is used. The words that this textism category targets (*you, are, see, etc.*) are very frequently used words. This means that even if only one textism form is used, there is a low OFL, since a majority of the time it is not being used. Instead, the standard spelling is used. This also explains why all non-singleton English forms are low usage. However, since there are still multiple contexts with a low-usage optimal form, rules can still be created to describe how the textism forms are generated. These rules are shown in Figure 10.

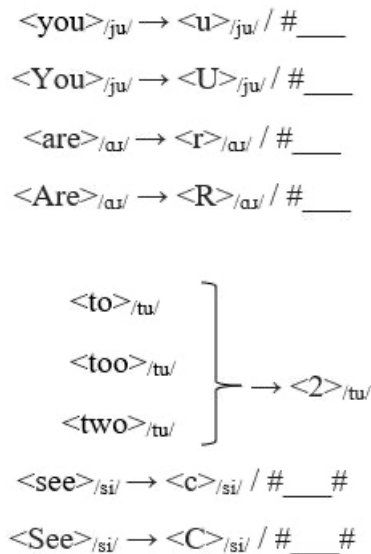


Figure 10. Rules for letter/number homophone

The subscripts represent the phonemic correspondences to the graphemic representations. An interesting observation is that for letter homophones, capitalization is preserved. For number homophones, capitalization is ignored, since there is no capitalization for numerical graphemes. These rules, especially for 2, can replace parts of words as well, such as *2night*, which is further evidence in favor of this category being rule-based, as opposed to there being numerical lexical variants of the English forms.

Missed Punctuation

Missed punctuation accounts for 30.7% of the textisms collected, with 514 instances of use. Missed punctuation is when punctuation that is normally required is omitted from a sentence. Missed punctuation only has one context of use (punctuation marks) and only one form that is produced by it (omission of punctuation marks). Therefore, its OFL defaults to 1.0 since all forms produced are optimal forms. Usage for missed punctuation is 447, meaning that missed punctuation is used a vast majority of the time instead of using proper punctuation. The rules for missed punctuation are shown in Figure 11.

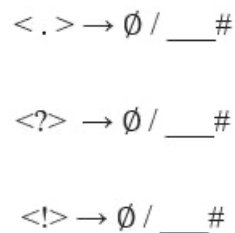


Figure 11. Missed punctuation rules

Missed punctuation serves a very important pragmatic role in texting. The lack of punctuation indicates a neutral tone when texting, while using a period can indicate a harsh or passive-aggressive tone (Shim, 2016). This harshness applies only to periods, which explains why the period is more likely to be removed than other punctuation marks.

Contraction

Contractions made up 3.7% of the total textisms, with 62 instances of use. Contraction is when letters, mainly vowels, are removed from the middle of the word. Table 10 shows the OFL and usage for all contexts and non-singleton contexts for contractions. Figure 12 plots these forms by OFL and usage.

	All Contexts (30)		Non-Singleton Contexts (10)	
	High OFL (≥ 0.5)	Low OFL (< 0.5)	High OFL (≥ 0.5)	Low OFL (< 0.5)
High Usage (> 0)	10 (33.3%)	3 (10%)	3 (30%)	3 (30%)
Low Usage (≤ 0)	8 (26.7%)	9 (30%)	1 (10%)	3 (30%)

Table 10. OFL and usage for singleton and non-singleton English forms of contraction

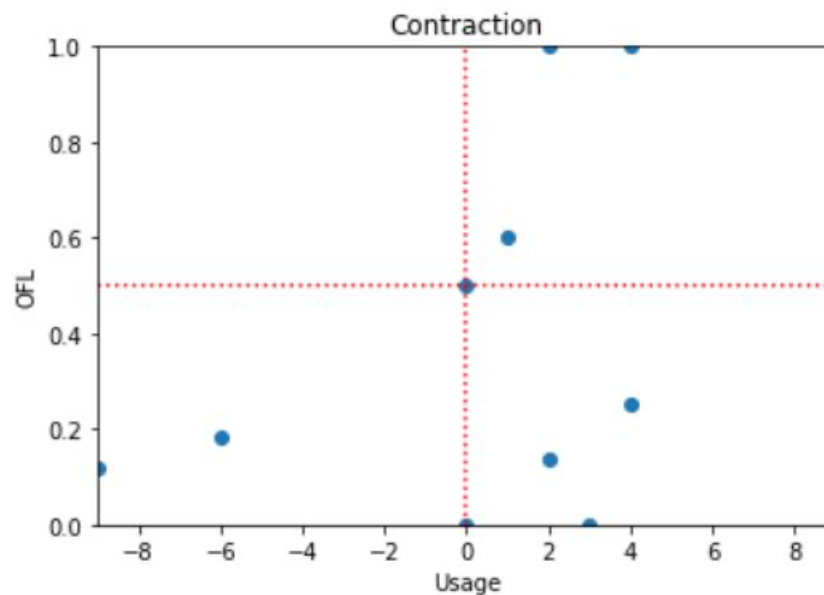


Figure 12. OFL and usage for non-singleton English forms of contraction

Table 10 shows that 33.3% of English forms (30% for non-singleton English form) have an optimal form. One factor that may be contributing to the low number of optimal forms is the two different types of contraction discussed in Kumar (2012). The first type of contraction is omission of all graphemes representing a vowel sound minus initial sounds, and the second type is the omission

of all graphemes except those representing consonant sounds. These two types of contraction may have led to more than one candidate form being accepted. Figure 13 demonstrates this by showing the textism distribution for the English form *Sorry*.

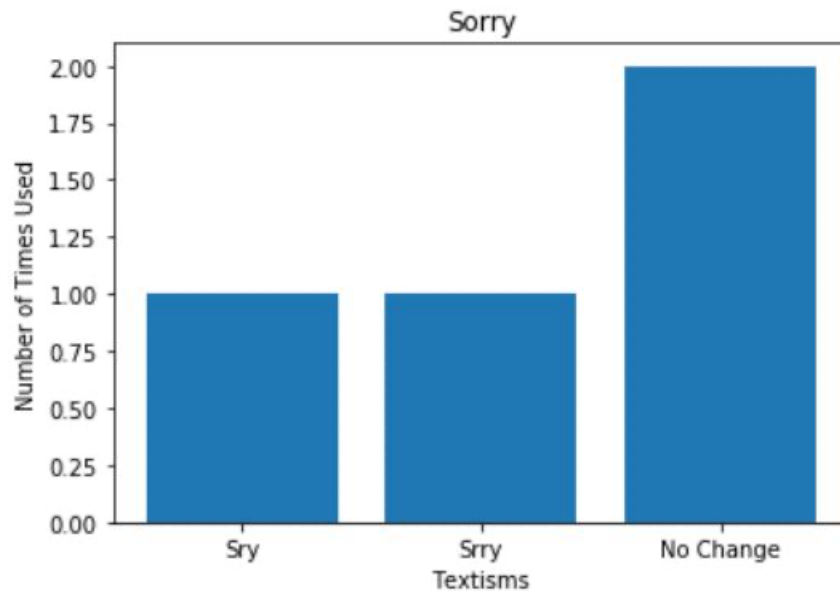


Figure 13. Textism distribution for English form "Sorry"

The textism *Srry* represents the first type of contraction, with only omission of the <o>. In this case, <y> is not considered a vowel, despite having the vowel sound /i/. The other textism, *Sry*, represents the second type, with omission of the <o> and an <r>, since only one <r> represents the consonant sound /ɹ/. This means that the English form *Sorry* for the category of contraction has an OFL of 0.0 and a usage of 0, despite both types outputting well-formed textisms. The rules for type one of contraction are formalized in Figure 14.

$$\langle V \rangle \rightarrow \emptyset / \langle G \rangle _$$

Figure 14. Rules for contraction type 1

<V> represents any vowel grapheme (<A>, <E>, <I>, <O>, <U>). This creates forms such as *abt* from the context *about*. The rules for type two are shown in Figure 15

$$\langle C_1 C_1 \rangle \rightarrow \langle C_1 \rangle$$

$$\langle V \rangle \rightarrow \emptyset$$

Figure 15. Rules for contraction type 2

<C₁C₁> represents two consecutive consonants that are the same grapheme (such as <rr> → <r> in *Sorry*). This notation differs from <C_{/p}/C_{/p}>, as this notation refers to a consonant grapheme

that corresponds to a certain phoneme p , while the notation used for this rule denotes two of the same grapheme used. This creates textism forms such as *ltrly* from the English form *literally*. There seems to be a favoring of retaining final letters, especially in the case of <y> in *ltrly*. However, this does not hold for other vowels in final positions, such as <e> in *msg*. In the case of *Sorry*, despite <y> corresponding to the vowel sound /i/, <y> does not act as a vowel grapheme.

Phonetic Spelling

Phonetic spellings accounted for 1.9% of the total textisms, with 32 instances of use. Phonetic spelling is when a word is respelled the way that it sounds. Table 11 shows the OFL and usage of all contexts and non-singleton contexts.

	All Contexts (18)		Non-Singleton Contexts (2)	
	High OFL (≥ 0.5)	Low OFL (< 0.5)	High OFL (≥ 0.5)	Low OFL (< 0.5)
High Usage (> 0)	10 (55.6%)	0	0	0
Low Usage (≤ 0)	3 (16.7%)	5 (27.7%)	0	2 (100%)

Table 11. OFL and usage for singleton and non-singleton English forms of phonetic spelling

The small corpus of phonetic spellings (only 32 examples) coupled with the broad scope of the category leads to very few contexts with an optimal form and only singleton English forms having an optimal textism form. Therefore, it is not possible to determine any specific rules due to the scarce data on the textism category and lack of non-singleton optimal forms.

Clippings

Clippings make up 0.8% of the total textisms, with 15 instances of use. Clipping is when the final letter of a word is omitted. Table 12 shows the OFL and usage of all contexts and non-singleton English forms for clippings.

	All Contexts (12)		Non-Singleton Contexts (1)	
	High OFL (≥ 0.5)	Low OFL (< 0.5)	High OFL (≥ 0.5)	Low OFL (< 0.5)
High Usage (> 0)	5 (41.7%)	0	0	0
Low Usage (≤ 0)	3 (25%)	4 (33.3%)	0	1 (100%)

Table 12. OFL and usage for singleton and non-singleton English forms of clippings

The most common form of clipping is g-clipping. Much like letter/number homophones, the English forms where g-clippings are applied (*-ing* suffixes) are very frequently used, causing the OFL and usage to drop when g-clipping is not used. However, unlike phonetic spelling,

conclusions can still be made for the rules of clipping due to its narrow scope of just omitting one grapheme in a set context. Figure 16 shows the rule.

$$\langle G \rangle \rightarrow \emptyset / _ \#$$

Figure 16. Rule for clipping

The $\langle G \rangle$ again denotes any grapheme. This rule deletes any final grapheme. Previous work on clippings has found multiple purposes for it, such as consonant cluster reduction (*wit* from *with*) and dialectal writing (g-clipping to represent African American Vernacular English, much like accent stylization) (Einstein, 2013).

Shortening

Shortenings accounted for 4.9% of the total textisms, with 83 instances of use. Shortening is when the end part of a word is removed. Shortenings are differentiated from clippings because shortenings omit more than one final grapheme, while clippings only omit the final grapheme. Table 13 shows the OFL and usage for all contexts and non-singleton English forms for shortenings. Figure 17 shows these forms plotted by OFL and usage.

	All Contexts (33)		Non-Singleton Contexts (15)	
	High OFL (≥ 0.5)	Low OFL (< 0.5)	High OFL (≥ 0.5)	Low OFL (< 0.5)
High Usage (> 0)	20 (60.6%)	3 (9.1%)	6 (40%)	3 (20%)
Low Usage (≤ 0)	6 (18.2%)	4 (12.1%)	3 (20%)	3 (20%)

Table 13. OFL and usage for singleton and non-singleton English forms of shortening

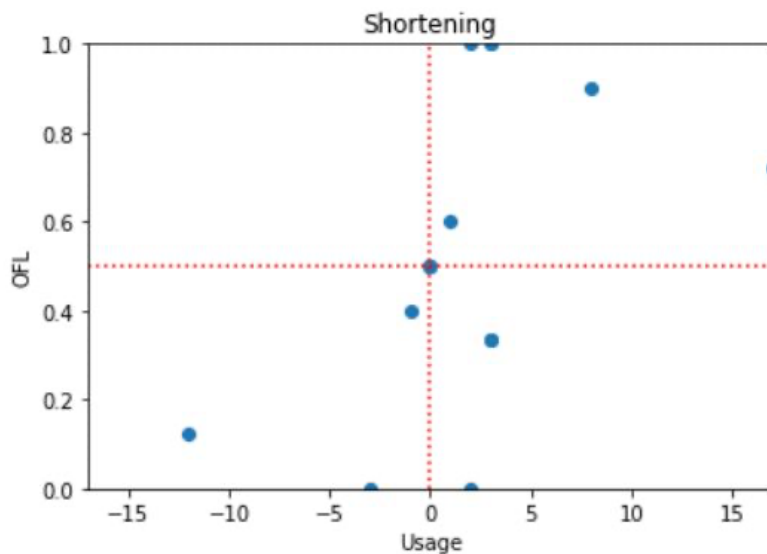


Figure 17. OFL and usage of non-singleton English forms of shortening

For shortening, 60.6% of English forms (40% of non-singleton English forms) have an optimal form with high usage, while only 12.1% of English forms (20% of non-singleton English forms) have no optimal form. McCulloch (2015) explains that shortenings are created by removing everything past the stressed syllable. The data collected in the present study agrees with this, with one exception. The English form *perfect* ['pɜr, fɪkt] when shortening is applied becomes *perf*. This cuts off everything after the stressed syllable except the onset of the next syllable, <f>. This exception is also applicable to the rest of the data collected. Therefore, this must be included in our rules for shortening, which are shown in Figure 18.

$$\sigma \rightarrow \emptyset / ' \sigma _ _$$

$$\sigma \rightarrow \sigma \langle C \rangle / ' \sigma _ _$$

Figure 18. Rules for shortening

σ represents a syllable, $'\sigma$ represents the stressed syllable, and $\sigma \langle C \rangle$ represents a consonant grapheme which is the onset of the syllable. The rules are also listed in order of application, having a counter-bleeding relationship, where if the rules were to be applied in the opposite order, the second rule would destroy the environment where the first rule would apply, leading to an ungrammatical textism. The first rule deletes any syllable two after the stress. This would change a word like *obviously* ['ab.viəs.li] to *obvious* ['ab.viəs], deleting the syllable two after the stress. The second rule then converts the syllable after the stress to just its consonant onset, changing *obvious* ['ab.viəs] to *obv* ['abv].

This process of shortening resembles a hypocoristic, where everything after the first syllable of a word is omitted, and a diminutive suffix is often added. These are seen in some Australian place names, such as *Tazzie* for *Tasmania*, or *Sevvo* for *Seven Hills* (Simpson, 2004). These examples follow a similar process as shortening, with *Sevvo* taking the onset of the next syllable. However, these hypocoristics are usually disyllabic, consisting of the original initial syllable and the diminutive suffix. Shortenings differ from hypocoristics by preserving everything before the stressed syllable. However, the two are similar in that their truncation is prosodically conditioned, specifically through their syllabic structure.

Initialism

Initialisms accounted for 24.3% of the total textisms, with 407 instances of use. Initialism is when a word or phrase is reduced to just the first letter of the word. Table 14 shows the OFL and usage for all contexts and non-singleton English forms of initialism. The forms are plotted by OFL and usage in Figure 17.

	All Contexts (101)		Non-Singleton Contexts (55)	
	High OFL (≥0.5)	Low OFL (<0.5)	High OFL (≥0.5)	Low OFL (<0.5)
High Usage (> 0)	82 (81.1%)	10 (9.9%)	41 (74.5%)	10 (18.2%)
Low Usage (≤ 0)	4 (4%)	5 (5%)	1 (1.8%)	3 (5.5%)

Table 14. OFL and usage for singleton and non-singleton English forms of initialism

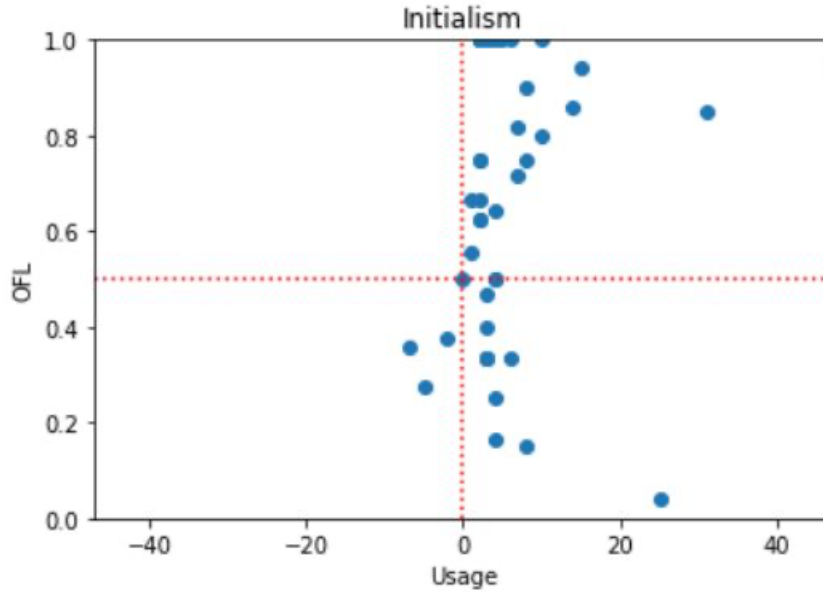


Figure 19. OFL and usage for non-singleton English forms of initialism

As Table 14 shows, 81.1% of English forms (74.5% of non-singleton English forms) have an optimal form with high usage, compared to only 5% of English forms (5.5% of non-singleton English forms) which have no optimal forms. The data shows two types of initialism: syllabic and phrasal. Syllabic initialisms are when a single word is reduced to just the first consonant of each syllable (such as *because* becoming *bc*). Syllabic initialism is described in Figure 20.

$$\sigma \rightarrow \sigma\langle C \rangle$$

Figure 20. Rule for syllabic initialism

$\sigma\langle C \rangle$ represents the first consonant of the onset of the syllable. This is applied to each syllable in the word to produce its syllabic initialism. Phrasal initialism is when the first letter of every word in a phrase is put together into a single word (such as *laugh out loud* becoming *lol*). The rules in Figure 21 would be applied to each word in the target phrase, listed in order of application:

$$\sigma \rightarrow \emptyset / \sigma _ _$$

$$\sigma \rightarrow \sigma \langle G \rangle / \# _ _$$

Figure 21. Rules for phrasal initialism

Much like shortenings, these rules have a counter-bleeding relationship; that is, they must apply in this order.

Onomatopoeia, Misspellings, and Semantically Unrecoverable

The textism categories of onomatopoeia and misspellings did not have enough data from which to draw conclusions, since all their contexts were singleton contexts. Onomatopoeia and misspellings accounted for 0.7% each of the total textisms, with each having 11 instances of use. Semantically unrecoverable texts made up 0.1% of the textisms, with only one instance of use. Rules cannot be derived for semantically unrecoverable textisms due to their incomprehensible nature. The single example of a semantically unrecoverable textism from the data was “*L.s.*”

Inter-Category Rule Ordering/Complex Textisms

Some of the textisms from the dataset did not fall into just one of the textism categories, but rather were affected by multiple textism categories. As these types of textisms have not been discussed in previous studies, they will be referred to as “complex textisms.” A complex textism is when an English form is affected by the rules of more than one textism category to produce its textism form. An example of this would be *g2g*, from the English form *got to go*. *g2g* is the result of both a letter/number homophone and initialism. Another example would be *lmaooooo*, from the English form *laugh my ass off*, being the result of both initialism and accent stylization.

One important observation about complex textisms is that there are rule-ordering relationships between the textism categories. In the case of *g2g*, there is a counter-bleeding relationship between letter/number homophone and initialism, where letter/number homophone must be applied first, or else initialism will destroy the environment where letter/number homophone can apply, resulting in *gtg*. Despite *gtg* being a grammatical textism, this application would still be a counter-bleeding relationship, since the application of initialism destroys the environment in which letter/number homophone can be applied. The same counter-bleeding relationship can be seen in *lmaooooo* between initialism and accent stylization, where initialism must be applied first or else the result of accent stylization will be destroyed. These relationships are preserved between complex textisms using the same categories, as seen in the example *lolllll*. The same counter-bleeding relationship between initialism and accent stylization is present.

This is a vital finding for demonstrating the existence of a grammar for Textese, as it shows that Textese is one cohesive set of rules and not just a collection of smaller grammars across the textism categories. While every combination of textism categories does not have an associated complex textism in the data to prove their rule-ordering, observations can still be made from the existing

data. Seven textism categories had complex textisms. Table 15 shows which orders these textism categories can be applied in to generate complex textisms. Orders were determined by finding complex textisms and applying the categories in different orders to test if they would yield the target textism.

		Category Applied First						
		Missed Capitalization	Missed Punctuation	Letter/Number Homophone	Shortening	Contraction	Initialism	Accent Stylization
Category Applied Second	Missed Capitalization		yes		ok		idk	imma
	Missed Punctuation	yes		u			lol	Imma
	Letter/Number Homophone		u				gtg, ily	u
	Shortening	ok						Addy, def
	Contraction						*hau	k
	Initialism	idk	lol	g2g, ilu		hbu		lmao
	Accent Stylization	imma		uuu	Addyyyy, defff	kk	lmaoooo	

Table 15. Order of application of textism categories for complex textisms

The cells in green represent complex textisms constructed in the correct rule order. The cells in red represent either regular textisms (due to one category application destroying the environment for another category) or an ungrammatical textism, marked with an asterisk. Textisms in red cells without an asterisk are still possible textisms under the proposed grammar, they just cannot become a complex textism via the application of another category’s rules, as the initial rule application destroyed the environment in which the second category could apply. An example of this is *defff*, a complex textism involving shortening and accent stylization. When shortening is applied first, *definitely* becomes *def*, and then (as seen in the Shortening column), accent stylization can be applied to create *defff*. However, if accent stylization is applied first, *definitely* becomes *defffinitely*, and then shortening (as seen in the Accent Stylization column) makes it *def*, destroying the effect of accent stylization. Therefore, these two categories are in a rule-ordering relationship. Categories not listed in the table had no complex textisms in the data. The only textism categories that do not seem to follow any rule-ordering are missed capitalization and missed punctuation, as seen by the presence of all correct examples in both the columns and rows of missed capitalization/punctuation. For complex textisms using either of these, missed capitalization/punctuation does not interfere with the application of any other category.

Discussion

Overview

It is possible to model Textese as a rule-based grammar. Generalizable, context-sensitive rules were written for eight out of the 11 textism categories (missed capitalization, accent stylization, letter/number homophone, missed punctuation, contraction, clippings, shortening, and initialism). The remaining three textism categories (phonetic spelling, onomatopoeia, and misspellings) were inconclusive as to whether rules could be written due to a lack of data for those categories. (Semantically unrecoverable textisms are not subject to rules, as they are unrecognizable and have no decipherable English form). However, previous research on one of the inconclusive categories, phonetic spelling, explains that phonetic respelling is based off already existing orthographic principles and spelling patterns (Tagg, 2009). This shows that there are rules to explain phonetic respelling, since the rules already exist through orthographic principles. Despite the lack of data for some categories in this study, there still seem to be trends that could be expanded into rules if more data were gathered (i.e., metathesis, or the swapping of two letters, in misspellings). This evidence is promising for future research on textism generation.

Textism generation also appears to be predictable, with relatively low variation. This was measured through the OFL metric, with English forms with a high OFL having a predictable optimal textism form. Naturalistic uses of textisms had far more optimal textism forms than textisms generated experimentally (such as in the translation task). This could have been due in part to the nature of the experimental tasks, with participants making textisms that they might not use naturally. However, the naturalistic textisms that were collected had relatively low variation. Table 7 shows that 63% of non-singleton English forms from the text message submissions have a high OFL. Along with this, only 12.5% of English forms have more than one accepted textism form used, showing that while variation does exist in Textese, it is more predictable than not, and therefore that textism generation is likely to be rule-based.

Confounding Factors

One confounding factor in this study was lexical replacement in the translation task. During translation, many participants opted to change a word to another word. For example, for sentence #6, *This restaurant has great food!*, the English form *has great food* in some translations was replaced with the term *bussin*, meaning something that is really good. In cases like this, the new word cannot be counted as a textism, as its relation to the context is solely semantic. While *bussin* is a textism, its English form is not *has great food*, and therefore it cannot be counted as such. This may have contributed to the low number of optimal textism forms in the translation task. Table 4 shows that of the 19 English forms, 10 of them did not have an optimal textism form. Lexical replacement contributed to this by lowering the number of times a textism was used for the English form, lowering both the OFL and usage measures.

Limitations

Another limitation of this study was the lack of negative input for Textese. The study initially set out to find what constitutes a grammatical textism and how said textisms were constructed. However, this was limited by the lack of data on what constitutes an ungrammatical textism. The grammaticality judgments provided a small number of ungrammatical textisms; however, there was not enough data on ungrammatical textisms on which to base any findings, only enough to confirm the existence of ungrammatical textisms.

This limitation is highlighted in the findings for initialism. The rules produced for initialism are based off the instances where initialism was used correctly, causing them to catch false positives such as *hagd* from the English form *have a good day*, which only had an acceptability rating of 7.76% in the grammaticality judgment task. Previous studies that have tried to determine what can and cannot be initialized concluded that initialisms are “unfathomable without prior knowledge of the referent or repeated use” (Shortis, 2007). While this might make it seem as though initialisms are determined lexically, Crystal (2008) finds evidence of productive initialisms, such as the initialism *imo*, meaning *in my opinion*. This initialism becomes productive once more words are added to the English form, such as *in my humble opinion (imho)* and *in my not so humble opinion (imnsho)*. This tendency is also seen with the lesser used *rofl*, meaning *rolling on the floor laughing*, being expanded to *rolling on the floor laughing out loud (roflol)* and *rolling on the floor laughing my ass off (roflmao)*. Since these initialisms show evidence of being productive, it is more likely that initialisms are generated through a rule-based grammar than generated lexically. Without more data on what constitutes an ungrammatical initialism, however, this is the closest we can approximate the rules of initialism.

Areas for Future Research

This study is a preliminary attempt at showing that Textese has a grammar, and further research is needed. While the corpus for the present study was not small (1,669 textisms), some categories lacked sufficient data for significant generalization. For future studies, a larger corpus of textisms is vital for drawing generalizations about textism generation.

Another important area for future research would be to investigate variation in Textese. Specifically, research should investigate if the same speaker uses different textisms for the same English form. If so, it would be evidence for variation in Textese. The geographic location of the speaker should also be considered to see if textism usage differs by dialect, building upon work such as Tateman (2015).

Conclusion

There is a long-standing view among language prescriptivists that Textese is a corruption of the English language, a “bleak, sad shorthand” (Sutherland, 2002) that augurs “the death of English” (Thurlow & Brown, 2003). These sentiments view Textese as non-linguistic and parasitic on

language. In contrast, the present study demonstrates that Textese obeys linguistic principles, and thus can be considered a language variant in its own right.

The study set out to investigate if textisms were created by linguistic principles that can be generalized through context-sensitive rules. To explore this proposition, 103 participants were given three tasks: a translation task, a grammaticality acceptability judgment task, and a text message submission task, and the textisms produced were classified into categories. Roughly three-quarters of the textism categories were found to have generalizable rules that explain how textisms are generated, while the remaining categories had insufficient data upon which to base any strong conclusions. Moreover, inter-category rule-ordering relationships were found for complex textisms, showing how the different textism categories interact with each other as part of a cohesive grammar, as opposed to a set of discrete grammars. Finally, textism generation was found to be predictable, with 63% of English forms yielding a single optimal textism form. This is compared to the 12.5% of English forms that had multiple optimal textism forms and were therefore less predictable. With a majority of textism categories having generalizable rules, no categories inconsistent with having rules, and a relatively low rate of variation, the evidence is compelling that textisms are generated through a rule-based grammar.

Acknowledgments

An earlier version of this paper was submitted as a bachelor's thesis at the University of Illinois Urbana-Champaign. My most sincere gratitude goes to my advisors on this project, Dr. Ryan Keith Shosted and Dr. Aida Talić. I would also like to thank Dr. Susan Herring and the two anonymous peer reviewers at *Language@Internet* for helping shape this paper into its final product. Finally, I would like to thank all of my participants on this project.

Note

1. The typographical conventions used in this article for conveying sounds and writing are as follows: \diamond represents graphemes, units of writing such as letters. [] represents phones, spoken units of sound. // represents phonemes, or mental representations of speech sounds, much like a phone. Phonemes are used for describing relationships between sounds and graphemes, while phones are used for describing actual speech.

References

- Chen, S. (2021). *How Do I Do This Asdfjkl – An investigation into the syntax of keysmashing*. Unpublished thesis, University of Illinois Urbana-Champaign.
- Coe, J. E., & Oakhill, J. V. (2011). 'txtN is ez fu no h2 rd': The relation between reading ability and text-messaging behaviour. *Journal of Computer Assisted Learning*, 27(1), 4-17.
- Crystal, D. (2009). *Txtng: The gr8 db8*. OUP Oxford.

- De Jonge, S., & Kemp, N. (2012). Text-message abbreviations and language skills in high school and university students. *Journal of Research in Reading*, 35(1), 49-68.
- Drouin, M., & Driver, B. (2012). Texting, textese and literacy abilities: A naturalistic study. *Journal of Research in Reading*, 37(3), 250-267.
- Eisenstein, J. (2013). Phonological factors in social media writing. *Proceedings of the Workshop on Language Analysis in Social Media* (pp. 11–19). The Association for Computational Linguistics.
- Halle, M. (1962). Phonology in generative grammar. *Word*, 18(1-3), 54-72.
- Kemp, N. (2010). Texting versus txtng: Reading and writing text messages, and links with other linguistic skills. *Writing Systems Research*, 2(1), 53–71. doi:10.1093/wsr/ws002
- Kemp, N., & Bushnell, C. (2011). Children's text messaging: Abbreviations, input methods and links with literacy. *Journal of Computer Assisted Learning*, 27(1), 18–27. <https://doi.org/10.1111/j.1365-2729.2010.00400.x>
- Kemp, N., & Clayton, J. (2016). University students vary their use of textese in digital messages to suit the recipient. *Journal of Research in Reading*, 40. <https://doi.org/10.1111/1467-9817.12074>
- Kent, S., & Johnson, G. (2012). Differences in the linguistic features of text messages sent with an alphanumeric multi-press keypad mobile phone versus a full keypad touchscreen smartphone. *Scottish Journal of Arts, Social Sciences and Scientific Studies*, 7(1), 50-67.
- Kul, M. (2007). Phonology in text messages. *Poznań Studies in Contemporary Linguistics*, 43(2), 43-57.
- Kumar, N. (2012). A linguistic study of abbreviations in SMS. *Language in India*, 12(6). <http://www.languageinindia.com/june2012/naveenabbreviations.html>
- Ling, R. (2010). Texting as a life phase medium. *Journal of Computer-Mediated Communication*, 15(2), 277–292. <https://doi.org/10.1111/j.1083-6101.2010.01520.x>
- Lyddy, F., Farina, F., Hanney, J., Farrell, L., & Kelly O'Neill, N. (2014). An analysis of language in university students' text messages. *Journal of Computer-Mediated Communication*, 19(3), 546-561.
- Mackey, A., & Gass, S. (2005). *Second language research methodology and design*. Lawrence Erlbaum Associates.
- McCulloch, G. (2015, October 16). *Abbrevs are def totes legit. but how do they get their spelling?* *Mental Floss*. <https://www.mentalfloss.com/article/69772/why-are-your-fav-abbrevs-totes-legit-hard-spell>
- McCulloch, G. (2020). *Because internet: Understanding the new rules of language*. Riverhead Books.
- Shim, M. (2016). *Is it really “fine”?* *An analysis of the paralinguistic function of punctuation in text messages*. Scripps Senior Theses. https://scholarship.claremont.edu/scripps_theses/831

- Simpson, J. (2004). Hypocoristics in Australian English. In K. Burridge, B. Kortmann, & E. W. Schneider (Eds.), *Varieties of English*, 3, 398-415.
- Shortis, T. (2007). *Gr8 txtpeceptions*. English Drama Media.
- Sutherland, J. (2002, November 11). Cn u txt? John Sutherland asks what texting is doing to the English language – and finds it all a bit :- *The Guardian*.
<https://www.theguardian.com/technology/2002/nov/11/mobilephones2>
- 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.
- Tagg, C. (2009). *A corpus linguistic study of SMS texting*. Doctoral dissertation, University of Birmingham.
- Tatman, R. (2015). #go awn: Sociophonetic variation in variant spellings on Twitter. In: S. Onosson & M. Huijsmans (Eds.), *Working Papers of the Linguistics Circle 25.2: Proceedings of the 31st annual North West Linguistics Conference* (pp. 97–108).
- Tatman, R. (2016). ‘I’m a spawts guay’: Comparing the use of sociophonetic variables in speech and Twitter. *Selected Papers from NWAV 44*, 22 (2), 161-170.
- Taylor, A. S. (2005). An SMS history. In L. Hamill & A. Lasen (Eds.), *Mobile world past, present, and future*. Springer.
- Thurlow, C., & Brown, A. (2003). Generation Txt? The sociolinguistics of young people’s text-messaging. *Discourse Analysis Online*, 1(1), 30.

Biographical Note

M Angel [mangel1@uchicago.edu] is a master’s student in linguistics at the University of Chicago. His research focuses on linguistic approaches to computer-mediated communication, specifically phonological approaches.