2018

PRAGMATIC FUNCTIONALITY OF PUNCTUATION ON TWITTER

Elizabeth M. Wright
University of Kentucky, ewr225@g.uky.edu
Digital Object Identifier: https://doi.org/10.13023/etd.2018.337

STUDENT AGREEMENT:

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained needed written permission statement(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine) which will be submitted to UKnowledge as Additional File.

I hereby grant to The University of Kentucky and its agents the irrevocable, non-exclusive, and royalty-free license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless an embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

REVIEW, APPROVAL AND ACCEPTANCE

The document mentioned above has been reviewed and accepted by the student's advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student's thesis including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Elizabeth M. Wright, Student

Dr. Mark Richard Lauersdorf, Major Professor

Dr. Edward R. Barrett, Director of Graduate Studies

PRAGMATIC FUNCTIONALITY OF PUNCTUATION ON TWITTER

THESIS

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Arts in the College of Arts and Sciences at the University of Kentucky

By

Elizabeth M. Wright

Lexington, Kentucky

Director: Dr. Mark Richard Lauersdorf, Professor of Linguistics

Lexington, Kentucky

2018

ABSTRACT OF THESIS

PRAGMATIC FUNCTIONALITY OF PUNCTUATION ON TWITTER

This work presents an analysis of punctuation use in computer-mediated communication (CMC); in particular, the present study aims to describe the pragmatic functions of nonstandard punctuation on Twitter, providing a corpus-driven overview of the distribution and frequency of nonstandard punctuation use, and an analysis of sampled tweets at the individual tweet level to estimate noise levels in the overall corpus. A survey was also conducted which aimed to identify user understanding of the affective content of nonstandard punctuation strings and to identify any possible effects of character repetition. Survey results indicate that linguistic content was the strongest indicator of affective understanding, type of punctuation (i.e., ?, !, and combinations thereof) was a weaker indicator of some affective content, and repetition was not found to be significant. The study argues that certain string types, possibly defined by punctuation type and not count, have large indexical fields of pragmatic meaning available to them, which are bounded by context. In light of these observations, the study also proposes distinctions/categories of punctuation strings and their associated pragmatic meanings.

KEYWORDS: Linguistics; Computer-mediated communication; Pragmatics; Punctuation; Twitter

Elizabeth Wright

July 25, 2018

PRAGMATIC FUNCTIONALITY OF PUNCTUATION ON TWITTER

By

Elizabeth M. Wright

Dr. Mark Richard Lauersdorf
Director of Thesis

Dr. Edward R. Barrett
Director of Graduate Studies

July 25, 2018
Date

This thesis is dedicated to my mother, who life did not let see my journey or scholarship, and to my father, who has supported me without hesitation in my endeavors.

ACKNOWLEDGEMENTS

To the continued support and guidance of my committee – Dr. Mark Richard Lauersdorf, Dr. Jennifer Cramer, and Dr. Kevin McGowan – I extend my sincerest thanks; your input was invaluable in shaping the questions pursued here. To my committee chair, this thesis would not have been possible without the additional time and effort you extended to me. Many thanks to my friends and colleagues, who provided emotional support and feedback as this project grew and took shape. To Brandon Jent, Matt Bresnahan, and Kelly Wright, who have served as pillars for me in my time at UK, I could not have pursued this work to my fullest extent without you. To my father, thank you for your continued support of my scholarship and life goals.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. Introduction

Up until recently, most linguistic studies of variation have focused on language in non-digital spaces, be it spoken or written in nature[1]. However, language used in digital spaces – for example, email, online chats and messengers, and text messaging – is semantically and pragmatically rich, and displays productive use of linguistic knowledge as well as orthographic and non-linguistic features to augment communication. In addition, use or non-use of orthographic conventions and other features (e.g., emojis and GIFs), and register conventions (e.g., formality of email versus text messaging) differ from platform to platform (Herring and Androutsopoulos 2015); although all digital language, the forms these varieties take vary widely, showing large flexibility and variation in use of orthographic and other features to convey linguistic meaning.

The variation present in digital language, or computer-mediated communication (CMC), is of interest to linguists because it shows parallels to both spoken and written language; there exists a general sentiment in the literature that "the language found in computer-mediated discourse does not strictly belong to traditional definitions of either writing or speech," (Ong 2011:212), pointing towards CMC as a highly innovative and adaptable register of writing that is able to create rich linguistic signals. CMC must necessarily position itself in relation to many prescribed or *de facto* orthographic norms; however, much of the orthographic variation observed in CMC functions to convey information also conveyed suprasegmentally or paralinguistically (i.e., prosody, facial expressions and body positioning) in speech. It also shows feature use that is not present in written norms or is not (obviously) paralleled by features in speech (Vandergriff

[1] Signed varieties have also been neglected in this area of research.

2013)[2]. Walther (1992) formalizes these observations, positing the Social Information Processing (SIP) theory, which proposes that CMC users employ textual resources and CMC cues[3] to convey and embed socio-emotional meaning; in addition, SIP theory proposes the motivations and mechanisms necessary for a CMC cue to form.

One such textual CMC resource used to convey pragmatic and suprasegmental information is punctuation, the general focus of this study. Punctuation, such as exclamation points and question marks, often conveys pragmatic information, particularly intensification of part or all of the linguistic signal (see Jackson 2016 for an overview of intensification and punctuation as an intensifier). As such, these pragmatic properties are already intrinsic to the characters, and are easily and often manipulated in CMC in order to augment the pragmatic signal. Previous research on punctuation will be discussed in more depth in §2.2.

At its core, this study is driven by three central questions:

1. What pragmatic functions does punctuation serve in CMC, particularly within the Twitter speech community?
2. What punctuation strings have pragmatic functionality (if any)?

[2] The present study distinguishes paralinguistic cues – suprasegmental information such as pitch, and nonverbal elements of communication such as body language – from CMC cues – orthographic or visual cues present in CMC that can, but do not necessarily, communicate similar information as paralinguistic cues. Much previous work seeks parallels or connections between paralinguistic and CMC cues (e.g. Kalman and Gergle 2010, Cho 2010, Lin 2016), and Schandorf (2012) argues that CMC cues are gestural in nature; however, the present study, in line with previous work such as Vandergriff (2013), assumes that paralinguistic and CMC cues are not inherently connected or reliant on one another, though they can be understood in relation to each other.

[3] A CMC cue is defined here as a feature of CMC that can convey socio-emotional information. CMC cues can be orthographic in nature, such as use of capitalization, punctuation, and spacing. CMC cues can also be less concrete; interaction with the norms of a platform – adherence or rejection to register, for example – can signal social information, and can thus function as a platform- or community-specific CMC cue.

3. What range of pragmatic meanings do these strings encapsulate?

Such functionalities and their ranges must be (at least roughly) identified before more global questions can be pursued regarding the function of punctuation in CMC and its interaction with other linguistic and paralinguistic elements. Given the current state of research more broadly, the breadth and depth of pragmatic functionality of punctuation in CMC is unclear, though prior work has identified numerous pragmatic functions of other CMC cues; an experimentally vetted framework through which to understand repetition of paralinguistic CMC cues (e.g., punctuation, emojis, reaction images) is also absent[4]. The present work clarifies what functions orthographic features can adopt, and cataloging the observed pragmatic functions could help to illuminate the possible range of functions available to nonstandard punctuation. Nonstandard punctuation is used here to mean any punctuation string greater than a single character (e.g., ??), excluding ellipses.

Moving forward, pertinent prior research on CMC, punctuation, and repetition will be covered in §2 in order to define the theoretical frameworks underpinning the present analyses. The general methodology of the study's two components and their results will then be discussed in §3; §3.1 discusses the corpus and subsequent frequency data to establish usage patterns, and then presents the analysis of sampled tweets from the corpus to assess general noisiness in the data; §3.2 discusses the survey design and results, which examined and discussed the factors determining pragmatic content in tweets, punctuation included. §4 presents a general discussion the findings of the corpus and survey, and the implications of the results taken together. §5 discusses limitations

[4] Jackson (2016) and Dresner and Herring (2010) offer promising frameworks, the former based in cognitive understanding of repetition and the latter looking through a lens of pragmatic functionality.

and future directions of the present study, and possible areas of investigation that could be helped by the research at hand.

2. Prior Work

2.1. An Overview of Computer-Mediated Communication

The present study analyzes linguistic data that was composed digitally; therefore, this study will be situated within the theoretical framework of computer-mediated communication. CMC encompasses a wide range of possible mediums and types of communication and is typically used as "a broad designator that encompasses multiple semiotic/linguistic modes… as well as technological interfaces" (Squires 2016:2). It includes text- and image-based modes of communication which take place through mobile phones, instant messaging (IM) interfaces, social media, etc. (Squires 2016).

CMC as a field began with studies scattered throughout the 1980s but only caught traction in the early 1990s amidst the sudden overload of digital communication and composition. The increase in usability (for example, through the development of user-friendly web browsers) and the expansion of the internet led to the advent of widely available synchronous and quasi-synchronous communication programs (Squires 2010) such as internet relay chats (IRCs), which gave users access to real-time communication[5]. The establishment of noticeable, archetypical CMC features such as "non-standard-typography, spelling, word-formation processes, and syntax" also further separated CMC from traditional written communication and speech (Herring and Androutsopoulos 2015: 131). Developments such as these precipitated the foundational question of whether CMC is more similar to speech or writing (Herring et al. 2013). CMC research also comments on spoken versus written similarity through the notion of synchronous and asynchronous

[5]It should be noted that other early platforms, such as email, could be used synchronously. However, IRCs were both by design and expectation synchronous, while this was not the understood norm with email.

communication. The development of real-time communication platforms (e.g., IRCs, chats, and messenger programs) removed the temporal barrier previously characteristic of written communication. CMC gave speakers[6] access to written linguistic resources and led to the genesis of new registers simultaneously[7].

Much early work viewed CMC as an impoverished register of communication[8], assuming that there existed communicative voids due to the lack of paralinguistic (e.g., facial expressions) and suprasegmental (e.g., prosody, pitch, intonation) cues[9]. These early studies (e.g., Carey 1980, Walther 1992, Baym 1995) focused mainly on how orthographic variation paralleled verbal or paralinguistic cues. A number of orthographic features have been clearly identified as carrying an interactional load, performing the interpersonal and pragmatic functions these early studies presumed necessary, including vocal spelling (Carey 1980, Lin 2016), letter repetition (Darics 2013, Kalman and Gergle 2014), and punctuation use (Gunraj et al. 2015, Squires 2012), among others. The manipulation of orthography is not a new phenomenon and has been researched outside of the digital space in mediums such as fiction and graffiti (Androutsopoulos 2000). However, in none of these other mediums is orthographic manipulation so regular and intrinsic as in CMC, which seems to show "loosen[ed] orthographic norms" (Darics 2013), and the motivation to fill perceived communicative voids left from speech.

However, CMC has grown out of this mold and moved on from questions concerning spoken versus written similarity to a more nuanced approach to variation.

---

[6] This thesis employs *user* and *speaker* interchangeably to refer to users of CMC.
[7] See work such as Cho (2010) and Ong (2011) for a discussion of variation and (a)synchronicity.
[8] This ideology was rooted in media richness theory, proposed by Daft and Lengel (1984). See Walther (1992) for an in-depth discussion of the theory's application to CMC.
[9] See Squires (2010) for further discussion of CMC's historically deterministic treatment of registers.

More recent CMC research has focused on identifying patterns of language change and variation (e.g., Bamman et al. 2014, Eisenstein et al. 2014), the effects of register and platform (e.g., Squires 2016a), discourse-level topics such as audience design (e.g., Iorio 2009, Androutsopoulos 2014, Pavalanathan and Eisenstein 2015) , and the interface thereof on CMC variation (e.g., Grouws et al. 2011). It is important to note that corpus-based and corpus-driven[10] studies in CMC have been utilized since the field's inception but are now appearing frequently and on much larger scales (e.g., Bamman et al. 2014, Eisensteing et al. 2014, Eisenstein 2015). Current research also seeks to identify constraints on variation in CMC, as the parameters and situational variables that control the use and diffusion of variation are still relatively unclear in digital spaces; users often cannot provide traditional sociodemographic variables (e.g., age, gender, ethnicity), avoid stating them, or falsify their content within digital communication platforms. Because of the difficulty in recovering and vetting available sociodemographic data that researchers face, there is a paucity of research on the sociolinguistic constraints in CMC, and it is unclear to what extent demographic variables influence variation in CMC[11]. Thus,

[10] The concept of a division between corpus-based and corpus-driven research began being discussed in works such as Sinclair (1991) and Tognini-Bonelli (2001). Here the distinction refers to corpus-based research as utilizing corpora as a methodology, but not concerned with data-driven questions, and as primed with questions before data has been seen. Corpus-driven research, on the other hand, is data-driven, and researchers do not form solid questions, but rather let the data drive the questions itself.

[11] Recent work rooted in network theory has shown follower networks and interest-based networks on Twitter to be the strongest predictors of lexical use (Eisenstein et al. 2014, Bamman et al. 2014, respectively). These networks differ from their non-digital counterparts in that they do not necessitate direct communication to any given node, that many of the edges are unidirectional in nature, many networks are asymmetrical (Squires 2012b), and that these networks are often interest- or topic-driven (Bamman et al. 2014). While characteristics such as age, gender, and race often mediate such interests, Bamman et al. (2014) identify differences between gender-based and interest-based networks, indicating that the networks along which CMC features and use spread are not simply digital reflections of those seen in physical space, and are in part mediated by particular social forces and to different degrees.

research must look language-internally to find constraints on variation; the present study aims, in part, to identify pragmatic constraints on nonstandard punctuation use.

Studies on variation in CMC began with the construction of classification systems to capture and categorize the variation emerging. In an early study of paralanguage in CMC, and in search for verbal correlates, Carey (1980) proposed five types of vocal spelling in CMC: lexical surrogates and vocal surrogates (e.g., onomatopoeic spellings of non-words such as *hehe*[12]), spatial manipulation (e.g., placement of letters to create an image, spaces indicating pauses), manipulation of grammatical features (such as punctuation and capitalization), and minus feature (the absence of particular features). Although this framework is meant to identify verbal correlation in CMC, Carey's framework is unable to fully account for more recent findings in which many CMC cues have expanded beyond indexing phonetic cues. In particular, a number of studies have looked at letter repetitions as means of non-verbal communication (Darics 2013, Kalman and Gergle 2014) and vocal spelling (Lin 2016, Kalman and Gergle 2014, Cho 2010).

For example, Darics (2013) examined IM data containing nonstandard letter repetition, collected from a workplace chat (Darics 2012). Darics finds that repeated letter strings, such as "allllllllllllllllooooooooooooottttttt" and "IIIIITTTTTTTT'SSSSSS THE WEEEEEKEND BAAAAAAAAAAABBBBBYYYYYYYYY!!!!!!!" (Darics 2013:144), can convey socio-emotional information (e.g., reluctance or excitement), evoke information from the auditory signal such as segment length or emphasis, and denote an informal register. Kalman and Gergle (2014) find that letter repetition often,

[12] See also Lin (2016) for an in-depth discussion of vocal spelling.

though not always, parallels mechanically feasible articulations, with 94% of the character repetitions present in the corpus being articulable. Those repetitions that were not articulable[13] were largely composed of letters representing stops, such as *t* or *d*. This indicates that letter repetition is likely used to parallel uses of phonetic lengthening present in speech such as filling pauses and emphasis; however, letter repetition allows for non-articulable sequences (e.g. sweeeeee**tttt**), showing that letter repetition in CMC does not rely on a phonetic parallel to encode additional meaning.

Baym (1995) tackled a broader classification question, naming five conditioning factors of language use through observation of an online topical discussion group: "external contexts – physical, cultural, and subcultural – in which CMC use is situated; the temporal structure of the group; the computer system infrastructure; the purpose of communication; and the characteristics of the group and its members" (Baym 1995, in Herring and Androutsopoulos 2015:130). Herring (2007) also proposed a more nuanced categorization of CMC types that combines medium and situational properties in order to classify discourse.

Because most linguistic innovation in CMC happens below the sentence level (Herring and Androutsopoulos 2015), orthographic variation is one of the more heavily-researched facets of CMC. Orthographic variables are typically most available to speaker appropriation and innovation, as they are a central, often necessary component of CMC, though they are not the only facets available to writers. Some work has also been done on

[13] Kalman and Gergle use the term articulable to mean able to be lengthened. It should be noted that repeated stop characters can be repeated in articulation, but this repetition does not achieve the same phonetic lengthening that non-stops can create. For example, while 'stoppp' could be verbalized by repeating /p/, this is not the same process as 'ssstop' or 'stoooop'.

orthographic variation as seen in punctuation use (e.g., Raclaw 2006, Squires 2012), letter repetition (e.g., Kalman and Gergle 2014), innovative spelling (e.g., Eisenstein et al. 2010), and phonetic correlation (e.g., Eisenstein 2015, Tatman 2012).

More recently, with the emergence of huge data availability (such as through the Twitter application programming interface, or API) and increases in computational power and tool accessibility, corpus-driven work has become the norm for CMC. Large corpora, even those with minimal metadata, allow for overarching variation trends and patterns to be viewed, and traced through a user population (e.g., Eisenstein et al. 2014).

The analyses undertaken here can be classified as computer-mediated discourse (CMD), which is a branch of CMC studies concerned with the discursive properties of CMC (Herring 2001, Herring and Androutsopoulos 2015). Herring and Androutsopoulos (2015) define computer-mediated discourse as "the communication produced when human beings interact with one another by transmitting messages via… any communication device" (127). The present work is concerned with the interaction of Twitter users[14]. CMD is distinguished from CMC more broadly through its focus on language use and use of discourse analysis as a primary method of analysis. While this separation is not acknowledged universally, it is important to note its existence, and to situate this work theoretically and methodologically not only within CMC, but also CMD.

[14] Although the platform used likely impacts the language composed due to affordances and constraints of each (e.g., availability of emoji and GIF keyboards on phones versus the ease of accessing less common punctuation on a traditional QWERTY keyboard), the present study will not touch on this. For a more detailed discussion of the impact of platforms, see Squires 2012a.

2.2. Punctuation

One orthographic resource available to CMC users to augment the linguistic signal, and the focus of the present study, is punctuation – ellipses, intensifiers, and dashes among others. Previous work has established that punctuation is involved with the communication of suprasegmental information (Ong 2011, Schandorf 2012, Vandergriff 2013, Lin 2016) and has been connected with emphatic expression and intensification (Schandorf 2012), particularly in conjunction with repetition (Jackson 2016). Schandorf (2012) identifies punctuation as a gesture which functions as a marker of emotional content and emphasis, namely through rhythmic structuring to convey suprasegmental information such as pitch and prosody. Dresner and Herring (2010) argue that punctuation, and emoticons with more intensity, are markers of illocutionary and perlocutionary acts and intentions, and as such can convey speaker emotional intent. Dresner and Herring's (2010) understanding of emoticons as conveyors of pragmatic meaning – not emotion – and their ability to do so without direct correlates to facial expression is important. They argue that CMC cues do not require a real-world emotive or gestural counterpart to convey specific affective information, and can evolve beyond one-to-one correlations with spoken phenomena to develop these affective meanings.

In addition to emphatic and suprasegmental information, punctuation can also convey social information about a user. Social information is often conveyed by the use or non-use of punctuation in particular digital settings; Gunraj et al. (2015) identify the period as a marker of both social and pragmatic information in text messaging, where use of a text-final period indicated insincerity of the author within the sample population; Squires (2012a) finds that use of apostrophes in a text message corpus correlates

significantly with gender, with women using apostrophes more than three times as often as men. Though the ellipsis can serve rhythmic functions, Raclaw (2006) argues that it can also indicate affective stance and in-group membership, serving as a marker of disagreement or distancing.

Previous work, such as that discussed above, shows that punctuation has the capability to be adopted as a CMC cue, and as such is able to convey social and emotional information *about speakers*. This research seeks to answer whether these established socio-emotional meanings are becoming more fine-grained or changing altogether in punctuation that already functions as a CMC cue.

2.3. Repetition

Central to this study is the notion of repetition, and in particular, the impacts of repetition on (pragmatic) meaning. A theoretical framework of repetition is required to interpret any results and to understand the mechanisms driving the repetition of punctuation and similar non-verbal features such as emojis. The theoretical notion of repetition and its effects are adopted from Jackson (2016)[15]. One key theoretical notion underpinning the current research is that repetition has emphatic and intensifying effects. The "emphatic nature of repetition" (Jackson 2016:34) has been well documented. It is clear that repetition of units, be it lexical, utterance, or morphological, allows speakers to highlight particular linguistic elements or concepts. Jackson speaks to the repetition of

[15] The full breadth of previous work on repetition is outside the scope of this work; however, Jackson's (2016) dissertation covers a wide range of linguistic and non-linguistic work on repetition of varying types, and discusses cognitive models for understanding the motivation and function of repetition, namely through the lens of Relevance Theory (Sperber and Wilson 1995).

intensifiers, the category to which question marks and exclamation points have been assigned here, arguing that the initial occurrence of an intensifier indicates the concept (e.g., shock, excitement, etc.), while further repetitions act as threshold markers for (often pragmatic) meaning. This notion of thresholds of meaning that are based in units of repetition is central to the structure of this thesis; Jackson provides theoretical grounding for the idea that different levels of incremental repetition can trigger different meanings or understandings[16]. Certain repetitions of punctuation could thus have different indexical fields (Eckert 2008) or carry different pragmatic meanings altogether. For example, any punctuation string containing a question mark may have a number of pragmatic meanings that it could possibly convey, likely in the semantic neighborhood of confusion (e.g., confusion, astonishment, bewilderment, etc.); the number of question marks in the string and the context in which it appears all work to narrow down which pragmatic meaning is conveyed by the string.

Jackson identifies the linguistic meaning of repetitions as non-propositional, highly contextual, and listener/speaker specific; in effect, she claims that the linguistic meaning of repetition lacks a conceptual interpretation, aligning repetition with characteristics of paralinguistic cues in CMC. This interpretation echoes Dresner and Herring's (2010) notion of emoticons as conveying illocutionary and perlocutionary meaning, rather than conveying any concrete concepts such as laughing or winking. While neither of these studies look at punctuation specifically, they discuss linguistic elements that serve paralinguistic functions, conveying vague pragmatic meaning that

[16] Jackson's interpretation of these threshold is largely cognitive in its motivations and explanation, thought this will not be discussed in the present thesis.

relies on context to be refined. This study assumes these conclusions are transferable to punctuation which has previously been identified as serving paralinguistic function (see §2.2.).

2.4. Social Media Data

With the advent of social media[17], enormous amounts of linguistic data have become publicly available to both users and researchers. Prior to this, much linguistic research on CMC was small scale, employing either publicly released data sets such as company email databases (e.g., Cho 2010, Kalman and Gergle 2014), forum data (e.g., Hardaker 2015), or chat logs often requested at a person-to-person level (e.g., Vandergriff 2013). However, now researchers have access to websites like Twitter and Reddit, which contain publicly available communications between users. Of course, not all social media is viable for linguistic research; websites like Facebook have strict privacy policies, and thus the amount of data that is public, intended for public consumption[18], and contains interactional data is quite slim and often raises ethical questions; platforms such as Instagram are mostly image-based, and contain minimal linguistic data; websites such as Tumblr are based around image and text-post sharing, and thus contain huge amounts of reduplicated data relative to original linguistic content.

[17] Social media platforms as we know them today were preceded by chat systems, such as IRC's, first invented in 1988 (Stenberg 2011) and systems like AOL instant messenger, released in 1997 (Petronzio 2012). The first social media platforms began to arise in the mid to late 90s. In 1997, a social networking website called Six Degrees was created, reputed to be the first social media website (Hale 2015). Myspace and Skype launched in 2003 (Crunchbase, Aamoth 2011), with Facebook launching soon after in 2004 (Carlson 2010).

[18] Many users do not realize their profiles or posts are public.

Central to the analysis of social media data is the question of audience design. The intended audience of an utterance has a hand in stylistic and other variation (Androutsopoulos 2014), therefore it is crucial to understand how unspecified versus specified audiences affect language variation. It is also crucial to be able to identify whether some audience is specified on a large scale, so that this variable can be accounted for in some way in corpus work.

While availability of data is of importance to CMC research, the availability of metadata is an essential factor to consider. Websites like Tumblr and Reddit are notorious for users providing extremely minimal and unreliable metadata, if any. Facebook contains perhaps the most metadata on individual users of any social media platform, but it can be assumed that the paucity of ethically sourced and publicly available data makes this platform less commonly utilized in linguistic research. Of all the social media platforms, Twitter seems to have the best balance of large-scale, viable linguistic data availability and cursory metadata attached to all users (e.g., user time zone and posting time are included in the metadata of each tweet posted). Because of the ease of data collection and the ability to collect large amounts of data, the present study used Twitter as the social media platform of focus.

2.5. Twitter

Twitter is a microblogging social media platform within which users can create posts, called *tweets*, up to 180 characters long[19]. As of January 2018, Twitter's userbase

[19] In September 2017, Twitter began the staggered release of an increased 280 character limit to users, and opened the new character limit to all users in November, 2017. It is worth noting that the majority of tweets

is skewed towards international users, with 79% of Twitter accounts based outside of the United States (Smith and Anderson 2018); within the U.S., approximately 24% of adults use Twitter with some regularity (Smith and Anderson 2018, York 2016). U.S. users are relatively evenly distributed within demographic categories, with 23% of women, 24% of men surveyed, and 24% of white, 26% of black, and 20% of Hispanic Americans surveyed reporting use of Twitter (Smith and Anderson 2018)[20]. As of 2016, Twitter's userbase was skewed towards younger adults, with 36% of users being 18-29 years old, 23% from 30-49 years old, 21% from 50-64 years old, and only 10% being 60 and above (York 2016).

Within the platform itself, users can retweet (in effect, retransmit) others' posts, create their own content, or both retweet and add onto another user's tweet. Tweets can contain other media as well, such as videos, pictures, and GIFs and reaction images, helping to augment the 180 character limit. User tweets and information are publicly available by default, although users can opt to set their profile, and thus all content produced and shared there, as private. All content produced by users who do not choose this option is available for fast, large scale data collection through the Twitter API; this makes Twitter a rich source of CMC data, although it is "demographically lean" (Iorio 2009), or impoverished in terms of available speaker demographic information (Squires 2016). Twitter social networks are often large and unidirectional, meaning that many

do not even approach either the previous or newly instated maximum; only 5% of tweets during the staggered test release of the 280 character limit exceeded 140 characters, and only 2% exceeded 190 (Rosen 2017). In addition, prior to the increased character limit, the majority of English tweets were only 34 characters – nowhere near the 180 character limit (Rosen and Ihara 2017).

[20] These percentages reflect engagement within demographic categories, and not overall percentages of the Twitter userbase, and thus does not necessarily reflect equal numbers of users from different races and ethnicities within the userbase.

tweets are intended to be seen by large audiences but not answered. This allows Twitter data to be analyzed linguistically without the surrounding context somewhat easier than other forms of CMC data, such as IM or email conversations[21,22].

As Twitter is a self-contained website, and users can communicate with each other and see the communication between other users, the present study considers Twitter to be a speech community, within which smaller communities of practice, or CoPs (Lave and Wenger 1991), exist through follower networks, and topical organization with hashtags. Twitter users converge in a shared location, and though they cannot realistically connect with all users, they share many nodes in their networks and have a shared purpose. It is through this theoretical framework that a shared pragmatic knowledge of punctuation and the possible diffusion through a larger network[23] is understood and argued for.

[21] This characteristic is rather important because when tweets are acquired, they come nearly entirely stripped of contextual information in a way that users would not see them presented.
[22] For a more detailed overview of Twitter mechanics and functionalities as a linguistic channel, see Gillen and Merchant (2013) and Squires (2012b).
[23] Here, the whole of Twitter, or the diffusion to other social media platforms based on a shared userbase.

3. Methods

In order to first identify which nonstandard punctuation strings are being used, a corpus-based inquiry was undertaken to provide a quantitative overview of punctuation use on Twitter. A perception survey was conducted to gauge affective perceptions of the same punctuation strings analyzed and quantified in the corpus. In addition, an analysis of 100 tweets from each punctuation class was conducted, to profile the corpus and evaluate noise levels in the data.

The corpus assembled for this study and its subsequent analysis serve mainly to provide a quantitative foundation on which to base the form and findings of the perception survey (§3.2.), and to illuminate the actual usage patterns and variation in these nonstandard punctuation strings on Twitter. In order to quantify and compare punctuation strings, categories had to be imposed on the data between which comparisons could be drawn. The punctuation marks covered in this study are question marks and exclamation points (categorized here as intensifiers) and commas and periods (categorized here as pauses). These categories were picked because of their productivity, both in standard and nonstandard uses, and the ability to compare within categories (e.g. pauses serve similar purposes, and as such periods and commas may follow similar constraints). It is within and between these categories that the current research will attempt to identify pragmatic meaning[24].

[24] Note that because pragmatic meaning, particularly in this case, is highly contextual, the categories imposed here draw boundaries that are far stronger than pragmatic and contextual meaning follow in actual usage. The current study, in addition to others (e.g., Jackson 2016), shows that pragmatic functionality is flexible and permeable across these boundaries. These boundaries are imposed only to quantify these phenomena into easily comparable categories. All results should be assumed as correlation, and not fixed effects of the punctuation strings they match to.

The categories delineated here vary slightly between intensifiers and pauses. Intensifiers are chunked/divided into one, two, and three or more instances (i.e., ! | !! | !!!+). Pauses are chunked into one, two, three, and four or more instances, (i.e., . | .. | ... | ....+).. Pauses are divided this way to account for the fact that three repetitions of a period already serve a prescribed function, an ellipsis, whereas there is no parallel feature of exclamation points and question marks, and as such, no reason to delineate such a category for them. In addition to viewing the data by degree of repetition, both same- and mixed-character repetition are covered here. Mixed-character repetition, strings such as *?!* or *!?!*, is only queried within broader function-based categories (here, pauses and intensifiers); any punctuation strings that might mix pause and intensification characters are not explored in the present work.

While these categories were established prior to any analysis[25], the following research provides empirical support for the boundaries drawn here, at least to the extent that they serve as rough pragmatic thresholds for users. To my knowledge, no previous work has investigated the pragmatic effects of repetition in punctuation specifically, although there is evidence for the gradation of intensity being correlated with repetition. Jackson (2016) discusses the correlation between affective impact and repetition of lexical units, arguing that there seems to be "a 'tipping point' where the markedness and effort of processing multiple adjectives encourages the hearer to adopt a different or additional processing strategy" (2016:228); there is a point at which continued repetition no longer has the same effect, being understood differently by the reader, and that units

[25] It should be noted that these categories were tested on an earlier pilot dataset whose results were presented at the Southeastern Conference on Linguistics in 2018.

of repetition leading up to this threshold feed into intensification of the same variety. For example, the repetition of 'very' in a sentence such as 'I'm very excited' only intensifies the writer's perceived excitement up to some (context dependent) number of repetitions, after which the repetitions begin to be understood as something other than intensification of excitement; perhaps sarcasm or an indication of informal register are signaled by these further repetitions.

3.1. The Corpus

In order to identify what punctuation strings are used within the context of Twitter (conceptualized here as a speech community) and provide an overview of usage patterns and frequency therein, a corpus of tweets was compiled. The corpus contains two separate datasets, collected four months apart; counts for individual tweets are shown below in **Error! Reference source not found.**.

Table 1. Tweet counts for each dataset and the overall corpus.

Dataset 1	**Dataset 2**	**Total**
n=4,236,232	n=3,419,496	n=7,655,728

The total tweet count is about 7.5 million (n=7,655,728). Each dataset represents a week of continuous data collection spanning from Monday at 00:00:00 to Friday at 23:59:59[26]. The first dataset contains roughly 4.2 million tweets (n=4,236,232), collected

[26] Previous work (Herdagdelen 2013) has shown that there are fluctuations in the user base and volume of tweet production on weekends, while weekdays stay relatively stable in terms of who is tweeting and how many tweets are published. For this reason, data was only collected during the most stable period, the work week, and weekends were eliminated in order to reduce known confounding variables interacting with the datasets

between January 15th and January 19th, 2018. The second dataset contains roughly 3.4 million tweets (n=3,419,496), collected between May 7th and 11th, 2018. Both were collected using the Twitter API[27], which was accessed via FireAnt (Anthony and Hardaker 2015), a software program that allows users to collect Twitter data and query it as a corpus. Two separate temporal slices were compiled in order to independently confirm the results from each, to have data across a wider diachronic and temporal window to identify whether these pragmatic meanings are (relatively) stable, and to account for any possible topical and temporal co-occurrences that could act as confounding variables (e.g., some current event that incites use of particular punctuation strings).

3.1.1. Quantitative Analysis and Results

Regular expressions were used to isolate and count the various standard and nonstandard punctuation strings. The regular expressions used are listed in full in Appendix A. The raw and normalized frequency counts of tweets containing each punctuation string are shown for the overall corpus and subcorpora in **Error! Reference source not found.** and **Error! Reference source not found.**. Tweets containing each identified punctuation string were output to individual .txt files, in effect creating smaller

[27] The Twitter API only provides access to publicly available tweets; anything published on private profiles or via direct messaging is not accessible. The Twitter API provides users with a random 1% of their search query, taken from the realtime feed of tweets being published. Thus, without any filters applied, the API would return 1% of all realtime tweets at random. This provides researchers with a premade random sample which, collected continuously over a longer span of time, allows compilation of a representative sample of Twitter as a whole. If filters are applied, the API still returns 1% of the search query, allowing for the creation of random samples based on a number of parameters (e.g., speech community via hashtags, time zone, etc.).

subcorpora organized by punctuation and repetition number; all subsequent results come from these subcorpora.

Table 2. Raw and normalized frequency counts for intensifiers.

	Dataset 1	**N per million**	**Dataset 2**	**N per million**	**Total n**	**N per million**
?	279,054	65,873	222,588	65,093	501,642	65,525
??	12,479	2,946	9,682	2,831	22,161	2,895
???	12,577	2,969	11,147	3,260	23,724	3,099
!	380,631	89,851	294,347	86,079	674,978	88,166
!!	50,552	11,933	42,920	12,552	93,472	12,209
!!!	52,383	12,365	44,822	13,108	97,205	12,697

Error! Reference source not found. and **Error! Reference source not found.** show both raw and normalized frequency counts. The raw counts are included simply to provide overall scale. The normalized tweets per million for each punctuation string are also broken down by subcorpora to verify internal consistency across the temporal slices, should there to be some confounding variable captured and reflected here. None of the normalized counts differ greatly between the two datasets, save for the per million count of triple comma strings.

Results for the intensifiers, shown in **Error! Reference source not found.**, show relatively frequent use of the standard variants: the single repetitions. The frequency of both exclamation points and question marks drop sharply with two or more repetitions, showing that these variants are significantly less common, but still have a strong presence in the corpus. In addition, normalized tweet counts for both intensifiers are similar for 2 and 3+ repetitions; however, the latter category encompasses more string types than does

the former, and these numbers do not provide a breakdown of where the bulk of tweets are within the 3+ categories.

Table 3. Raw and normalized frequency counts for pauses[28].

	Dataset 1	**N per million**	**Dataset 2**	**N per million**	**Total n**	**N per million**
..	57,437	13,559	44,722	13,079	102,159	13,344
…	163,853	38,679	124,709	36,470	288,562	37,692
….+	45,116	10,650	38,921	11,382	84,037	10,977
„	2,115	499	1,940	567	4,055	530
„„	771	182	1,285	376	2,056	269
„„„+	449	106	472	138	921	120

Similarly, **Error! Reference source not found.** shows the highest frequency counts in the standard cell for periods: three repetitions, or the ellipsis. This pattern does not hold for commas; in fact, commas show a steady decline in frequency inversely correlated with repetition number, and show up quite infrequently in the corpus overall. It is unclear from these data alone what is driving this nonstandard comma usage and whether or not it is meaningful. The differences in frequency counts between cells are small enough per million that many of these tweets could be typos, and they do not seem to simply be an alternate character choice for ellipses.

[28] It should be noted that single repetitions of both commas and periods were not included due to logistic issues. Because the syntax of JSON (JavaScript Object Notation), the notation used to encode the data in tweets, necessitates individual commas and periods at various places to delineate different blocks of information, regular expressions returned the entire corpus when searching for a single comma or period. Periods and commas were also present within links and embedded media, which show up within the body text of the tweets; because of this, attempts to focus the regular expressions to only the body text of the tweet failed, and returned huge amounts of noise.

3.1.2. Noise Evaluation and Results

Following the quantitative overview, an attempt to profile the corpus and provide some indication of overall noisiness of the dataset itself was undertaken. A window analysis was done; one hundred sequential tweets were selected from each punctuation category within the corpus. All tweets were selected from the newer dataset (see §3.1) so that results are temporally proximal to the data that most reflects current usage on Twitter[29].

Table 4. Total counts for spam and repeated tweets, tweets not penned in English, and advertisements.

	Spam	Repeats	Not English	Ad	Total Viable[30]
!	11	8	0	14	**86**
!!	11	9	1	10	**88**
!!!+	7	11	4	9	**88**
?	8	1	0	19	**91**
??	8	6	5	8	**86**
???+	5	16	0	1	**87**
..	2	11	3	4	**88**
…	34	33	0	5	**66**
….+	1	8	1	3	**94**
„	9	11	0	3	**89**
„„	4	36	0	0	**64**[31]
„„„+	4	16	1	0	**85**
?!+|!?+	12	4	1	10	**85**
.,+|,.+	23	22	4	3	**73**
Total	**139**	**192**	**20**	**89**	**1170/83.57%**

[29] All sampled tweets were taken from the exact middle of each subcorpus of punctuation type and repetition.

[30] This column represents the total number of viable tweets within each sample. Tweets were eliminated if they were categorized as spam, were repeated (the first instances were kept, but all subsequent instances were eliminated), or were not in English. Some tweets fulfilled more than one of these criteria, and so the total varies somewhat from the feature counts also reported in **Error! Reference source not found.**.

[31] The sample set for triple commas returned very low viable frequency counts. This sample set in particular contained very low diversity, as most of the sample was made up of three retweets. This could be because the punctuation string itself is quite low frequency, with only 2,056 instances in the entire corpus. Frequent retweets of popular tweets are diluted in the larger subsets, but with a low frequency string, there is little other content being produced to balance out the retweets.

Error! Reference source not found. shows the number of spam and repeated tweets of the 100 sampled per cell, the number that were not written in English, either in part or in full[32], and the number that appeared to be advertisements for companies or products. Totaling 1,400 tweets, this sample projects that about 83.6% of the corpus is composed of viable tweets. Most of the tweets marked as noise are high repetition retweets from spam accounts. Projected onto the corpus overall, around 1.2 million tweets are noise.

3.2. The Survey

While predictions can be made by the researcher as to the pragmatic function of nonstandard punctuation strings, these meanings are ultimately understood contextually by individuals. A shared understanding may bind a community of practice (here, likely a dense and multiplex network of followers), but a quantitative data-driven corpus analysis can only speak to frequency and distribution of forms; alone, it cannot answer questions of pragmatic meaning conveyed at the level of the utterance. In order to address this particular question, a survey was distributed through Qualtrics[33] to identify how a larger population understands the affective nuances of nonstandard punctuation, and which strings exemplify a shared understanding of such nuances. The survey required

[32] All tweets that appeared to be penned in another language, in part or in full, were counted as noise. The corpus is intended to contain only English tweets, and it is unclear from individual tweets whether or not the user is a native English speaker, or if English is a second language. In order to account for as many possible confounding variables, such tweets are marked as noise. However, tweets that appeared to discuss foreign references, such as celebrities, politicians, or places, were not excluded, as this was not grounds to assume any non-English language proficiency if the rest of the text was in English.

[33] Qualtrics is research platform that offers a number of data collection and analysis tools, including survey creation and hosting.

participants to rate ten tweets along eight emotional dimensions. Each question drew from a bank of stimuli, all with the same sentence but either three or four permutations of a single punctuation type each time. Thus, the linguistic content was controlled for, and the number of repetitions was the variable; this allowed for the affective content of each punctuation type and the effect of degrees of repetition to be identified.

Participants were shown 10 stimuli in total. They were first asked to indicate which sentence type (statement, question, or exclamation) the stimulus fit, and then were asked to rate the stimulus along a list of eight emotional dimensions: alarm, surprise, anger, excitement, confusion, annoyance, accusation, and offense[34]. Participants were asked to rate this emotional content along a slider which went from 0 (The author does not seem to be…) to 100 (The author definitely seems to be…)[35,36]. Following the 10 stimuli questions, participants were asked to provide demographic information regarding their age, race, social media use, and native English speaker status. The questions as they appeared to participants and the full list of possible stimuli are available in Appendix B.

[34] There is a definitive negative skew to these emotional states; while there are emotions with positive polarity, many of them did not match up with the function of intensifiers and felt forced or out of place. In order to avoid survey questions that felt unnatural, more negatively skewed emotions that related to the stimuli were selected in favor of balanced polarity.

[35] The emotions themselves were phrased as adjectives, and the questions asked whether the author, not the tweet, seemed to be angry, upset, excited, etc.

[36] A sliding scale was used in lieu of a Likert scale in order to encourage more decisive measurements (i.e. to not provide 'sort of agree' and 'sort of disagree' categories for answers), and for reasons of more straightforward statistical analysis. The sliding scale was only labelled in three places: far left- 0, strongly disagree, middle- 50, neither agree nor disagree, far right- 100, strongly agree. While participants could put the slider anywhere along the scale, any in between measurements were not labelled or identified as on a 5 point Likert scale.

Figure 1. Example stimulus. Stimuli only differed in the body text, and not the frame surrounding the body text.

The stimuli for each question consisted of a single tweet, created using a tweet generator, "Simiator" (Twitter Tweet Generator); the general appearance can be seen above in Figure 1. As it was optional when using the tweet generator, minimal metadata was included in the constructed tweets in order to simplify the stimuli; the embedding option, as well as time and date of publication, were not included[37]. No distinctive handle or username was used; rather, the account was named Anonymous and the handle was Anon, both of which are likely familiar terms for social media users[38]. Otherwise, the tweets shown in the survey reflect the appearance of a standard tweet.

The survey was distributed to two summer WRD[39] classrooms at the University of Kentucky as an extra credit option and on social media as well. WRD students had access to a link directly to the survey, and the instructors verbally explained the extra credit opportunity and presented a recruitment message composed by the researcher. The

[37] Removing the timestamps also served the added function of not temporally marking the stimuli. This allows for possible confounding temporal variables to avoided in the survey, such as how participants would interpret a tweet separated from them temporally.
[38] A distinctive handle and username were not used, as they could provide information regarding the author which could confound the results depending on what the identifiers were. In addition, participants may not all converge on the same demographic inferences; a handle may seem feminine to some participants, but not others. It is unclear whether this could affect results, and as such was eliminated as a possible confounding variable.
[39] The classes, both WRD 110 (WRD = Writing, Rhetoric, and Digital Media), are required introductory composition classes for students at the University of Kentucky.

recruitment message explained that the survey was minimal risk, and how long it would take. Those who came from social media saw the same recruitment message. Thirty-one results were obtained, eight of which were removed due to incompletion or noncompliance with survey instructions; twenty-three viable responses remained and are analyzed below. Of the respondents, thirteen self-reported as female, nine as male, and one as nonbinary. Eleven identified as White, ten as African American[40], and two as Asian. Fifteen participants were between the ages of 18 and 24, five were between ages

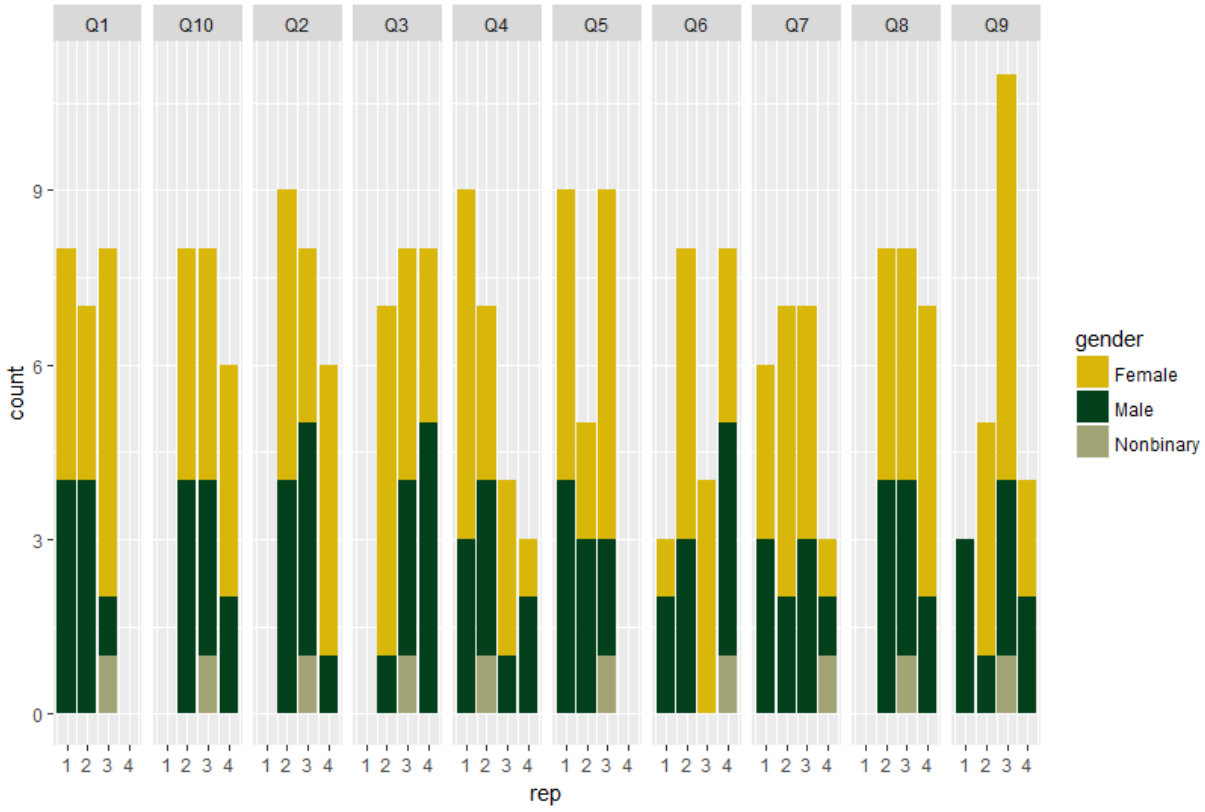


Figure 2. Distribution of responses across repetition number for each question, faceted by stimulus variety, and colored for self-reported gender[41].

[40] It is worth noting that one participant selected other and wrote Black as their race, although African American was provided. For the purposes of analysis, they were categorized as African American.
[41] Figure 2 and Figure 3 show the number of responses each stimulus received on the Y axis, classified by repetition number within the stimulus on the X axis. Repetition number refers here to the same strings

25 and 34, one was age 35 or above, and two declined to answer (although all had agreed via the consent form that they were above 18 years of age). Fourteen participants said that, among other platforms listed, they used Twitter on a regular basis; of those, five said Twitter was their most frequently used platform. The full report of demographic data broken down by participant can be found in Appendix C.

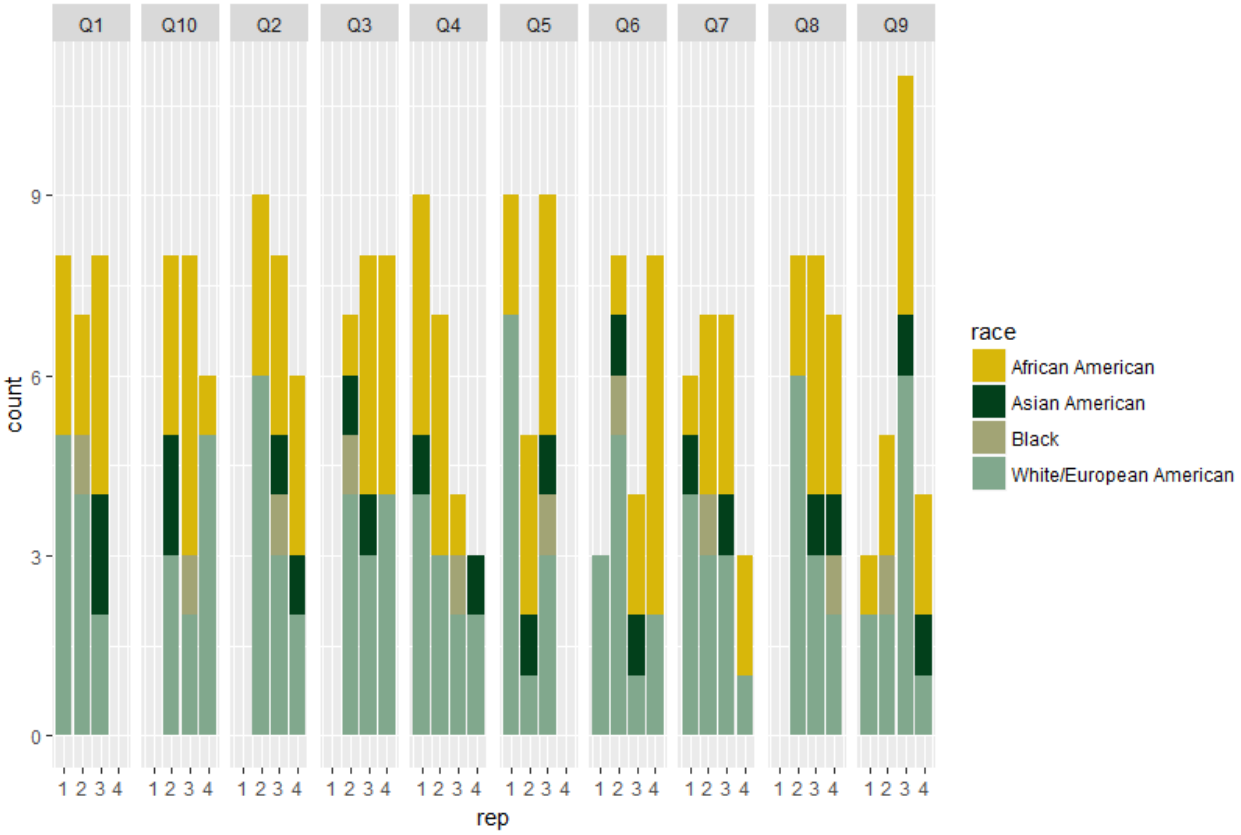


Figure 3. Distribution of responses across repetition number for each question, faceted by stimulus variety, and colored for self-reported race.

identified in the corpus analysis, with 1 being standard use, 2 being strings such as ?? and !!, and so on. These graphs are faceted by question number (here indicated by Q#) so that for each individual question (and thus, for only a single class of punctuation within each facet) the distribution of responses across the variable of repetition can be shown. The colors indicate demographic information provided by the participant.

Data was imported from Qualtrics, cleaned[42], and loaded into RStudio (RStudio Team 2016), where all subsequent data analysis was conducted. Figure 2 and Figure 3 show the total number of responses collected for each stimulus, with gender and race indicated respectively. Because the stimuli were drawn at random from a question bank, responses are not distributed evenly across variables, even within each question, and so it is important to visualize their distribution.

Before data analysis and results are discussed, it is important to note that the total response count (n=23) is quite low, especially for the survey design implemented here. This means that the number of responses in any given cell is relatively small, and thus the results discussed here may not hold up well given further data or in statistical analyses. As such, the quantitative results of this survey are likely mutable, and must be understood in relation to the other analyses undertaken here. The linear regressions run below were applied to very few datapoints. Because of the low number of participants, the results of the survey will be discussed in further detail beyond a purely quantitative analysis, as a qualitative analysis of the observed correlations is more helpful than a quantitative analysis in this case.

Mean values were obtained for each emotional attribute[43] (i.e. anger, confusion, etc.) by question number, repetition count, and type of punctuation using the aggregate

[42] The dataset output by Qualtrics was wide and needed to be converted into a long form in order to be read properly in R. Wide format data typically stores one datapoint (here, a response to a single question) per column, and rows indicate variables or characteristics associated with each response; long format data flips this, with each row containing a single response with variables and other associated information being indicated in the columns. In addition, there were many categories of information that were unnecessary for the data analysis undertaken here. Qualtrics records data such as response time, length of time spent on each question, and user location, among other things, that were not relevant or would not be analyzed in the present study.

[43] Here, emotional attribute and emotional dimension mean the same thing.

function of R[44]. Question number is included because it represents linguistic context of the variable punctuation strings, and as such represents the effect of semantic context. These averages were run through linear regressions using R's linear modeling function. Regressions were run for the intercept of each attribute's average scores and the punctuation type, number of repetitions, or question number (i.e., avg ~ type, avg ~ rep#, and avg ~ question#), for the intercept of punctuation type and number of repetitions (i.e., avg ~ type + rep#), and for the intercept of punctuation type, number of repetitions, and question number (i.e., avg ~ type + rep# + question#). Fit of the models was determined by the F-statistic given with each regression run; fit varied widely between models, even within each emotional attribute. Again, linear regressions were run based on attribute averages as they interacted with the other variables.

No single variable was significant across all regressions. Question number (and thus linguistic content accompanying the punctuation string) was significant in some, but not all, regressions and for some, but not all, emotional attributes in multiple regression models. For *alarm*, question number was significant for every question, and for *surprise* all questions save for Q9 were significant ($p \leq .05$ in all cases). For other attributes, only some question numbers were significant. For *anger*, only Q5 was significant ($p < .01$); for *excitement*, Q5, Q6, and Q9 were significant ($p \leq .05$); for *confusion*, Q5 and Q9 were significant ($p < .05$); for *annoyance*, Q5 was highly significant ($p < .001$) and Q8 was also

[44] A mistake was made in survey design, and responses to the sliders was not forced for completion of each question. Because of this, there were a large number of empty data columns where participants made no choice. In order to run averages, all N/A cells representing slider answers were replaced with values of 50, which indicated neither agree nor disagree on the sliders. It is assumed that, because participants did not interact with the sliders, they did not feel that emotion applied to the stimulus in any meaningful way and would have put the slider at or near 50 if a response was forced. This decision could have obscured the responses that actually reported 50s, or could have skewed the results away from significance.

significant ($p<.05$); for offense, Q7 and Q8 were significant ($p\leq.05$); no question numbers were significant for accusation. These values are all taken from the regression of attribute average ~ question number, repetition number, and punctuation type. Significance levels were the same in models only looking at interaction between attribute average and question number, indicating that additional intercepts did not measurably affect their significance.

The significance of question numbers refers to the effect of linguistic content on the assignment of values to the various emotions; this should reflect the understanding of emotional content in the stimulus. Because Q5, Q8, and Q9 were significant when other questions were not[45], they will be discussed further, and the single instance stimuli for each are shown below in Figure 4, Figure 5, and Figure 6, respectively. Q5 was a significant factor in the determination of annoyance, confusion, and anger; these are all emotions the text could convey, particularly at higher repetitions. It also was a significant factor in determining excitement, alarm, and surprise, though the latter two were significant for every question. Q8 was a significant factor in the determination of offense and annoyance in addition to surprise and alarm. It was not significant in the determination of other possible emotions, such as anger. Q9 was a significant factor in the determination of confusion, excitement, and alarm.

[45] Question nine was written in all caps, giving participants another cue to convey affective content. This stimulus was expected to be rated more extreme than other questions, as it has two markers of intensification, whose impact on the pragmatic content was expected to work synergistically, and did.

Figure 4. Stimulus from question five, single character variant.

Figure 5. Stimulus from question eight, single repetition variant.

Figure 6. Stimulus from question nine, single character variant.

Type of punctuation was significant in the determination of some emotional attributes. The regressions categorized mixed punctuation and same-character punctuation separately; as such these interactions were evaluated separately. Mixed punctuation strings were significant for determining alarm and confusion. Same-character punctuation strings were significant in determining confusion and excitement. Mixed and same-character punctuation strings only shared one indicative emotion, confusion, likely because both categories share the question mark.

Perhaps most importantly for the research question, the number of repetitions was not significant in any of the linear regressions run, either by itself or in conjunction with other variables. This indicates that the present data, at least as told by linear regressions, does not indicate any interaction of number of repetitions with affective content of a tweet. This directly rejects the central hypothesis and accepts the null hypothesis: repetition of punctuation has no effect on pragmatic content of a tweet. However, as we must be critical of the above results in favor of the central hypothesis, we must also be critical of these results as rejecting it. The categories that indicated number of repetitions were the least populated cells in terms of responses; n=23 in the cell for question number[46], as all responses included here completed the full survey. The responses were then split by punctuation type, which represented three categories; participants saw six same-character stimuli (three question mark strings, three exclamation point strings) and four mixed character strings. So, n=69 in cells for question marks and exclamation points, and n=92 for the mixed character cell (see Figure 2 and Figure 3 for the actual counts and distribution of responses across conditions, and **Error! Reference source not found.** below for a visualization of these numbers).

Table 5. Visualization of the number of data points obtained for each variable.

Variable	Number of Data Points
Question number	23
Type: !	69
Type: ?	69
Type: Mixed	92

[46] This is possibly why question number was found to be significant in many of the linear regressions run.

Repetition functioned differently from punctuation type and question number in that it was dependent on punctuation type; all same-character punctuation could be repeated between one and three times, with four questions allowing strings of four or more exclamation points or questions marks. Mixed character strings could only contain either two, three, or four plus characters (e.g., ?!, ?!?, ?!???), as they necessarily contain at least two characters. With four categories spread over ten questions, rather than three categories, response number per cell was quite low ($n \approx 57.5$) given the interactions of interest.

However, repetition still was not identified as a significant variable in any of the linear regressions run. It is possible that the effects repetition has on pragmatic meaning require more context – perhaps, conversational threads or user metadata – in order to be realized in the reader's comprehension of the linguistic signal. It is also possible that the subject pool was too diverse demographically in relation to the overall frequency count. Should the pragmatic meaning of punctuation shift between speech communities, or even individual CoPs, a subject pool diverse across race, gender, and age may have resulted in numerous pragmatic understandings being represented in the data, and subsequently washed out in analysis. A more robust response count and a more focused survey would likely provide enough responses for the effect of repetition – should there be one- to be visible in the statistical analysis.

4. Conclusions

Both prior research and the present study have shown that punctuation functions on Twitter as a CMC cue which transmits an author's illocutionary force, and like many other CMC cues, it does not necessarily retain original functions of the characters, having expanded into a flexible and productive pragmatic marker. In connecting the threads of analysis undertaken in the present thesis, we will revisit the central questions driving this research. First, what pragmatic functions does punctuation serve within the Twitter speech community? Both the current thesis and prior work indicate that context is paramount in defining and constraining the pragmatic function of (nonstandard) punctuation, particularly because punctuation is necessarily non-propositional[47]. For those survey questions where there was general consent on pragmatic content, the utterance itself was pragmatically rich; presented with a less accessible or understandable utterance, participants were unsure how to interpret the general emotional content of the stimulus. One participant noted in the final comment box provided that many emotions were hard to judge because there was no conversational context or user information. However, the survey provided no indication as to what constrained participants' understanding other than the linguistic context accompanying the punctuation string[48].

The second question asks which punctuation strings, if any, display pragmatic functionality. While the corpus indicates that some strings are used infrequently, it seems that nonstandard punctuation as a category is able to convey pragmatic meaning, given

[47] Some punctuation strings are able to function as discourse markers, and as such are able to stand alone. It is unclear whether these strings (such as a message containing only exclamation points) are truly non-propositional.

[48] As discussed in §3.2.1, demographic factors are likely a constraining factor in a reader's understanding of a punctuation strings' pragmatic content; however, the survey did not produce enough responses to tease out any correlations based on demographic information.

the correct contexts. Only one category of punctuation strings, mixed pauses, seems largely unproductive; all other categories of both mixed and same-character punctuation strings appear in the corpus and appear capable (to the researcher) of conveying pragmatic information. The survey did not address strings not already proven to be productive pragmatic markers.

The third question asks what range of pragmatic meanings pragmatically productive strings encapsulate. Again, linguistic context must be referenced here. Neither the corpus or survey results are able to identify with certainty boundaries of pragmatic function available to punctuation strings. Indeed, on the contrary, a wide range of pragmatic meanings were observed in the corpus, and participants identified a range of emotional content in controlled survey stimuli. Given the findings of this thesis, it can be concluded that nonstandard punctuation can convey a wide variety of pragmatic information, but no constraints can be placed with certainty on what that information is, or which punctuation can convey it.

5. Future Directions and Limitations

Left largely unanswered by the present thesis is the question of the pragmatic limits of nonstandard punctuation. The pragmatic content that particular punctuation strings are able to convey could not be defined here; future work should focus on the boundaries of pragmatic functionality of particular strings with the goal of identifying limits. The identification of pragmatic boundaries (or a lack thereof) would provide insight into how punctuation functions as a CMC cue. One topic not addressed here is the larger implication of punctuation as a CMC cue. Only four types of punctuation were addressed here – this presents the question of whether other punctuation marks are being used as CMC cues, and if so, what paralinguistic information are they conveying? Are there any unifying characteristics that punctuation-based CMC cues share, or are any connected only by an orthographic class?

The present study was also unable to provide concrete evidence on the effects that repetition of punctuation has on its pragmatic content. A more refined survey with a higher response count could isolate any possible pragmatic implications of repetition or confirm that the effects of repetition are too indeterminate, context specific, and speaker specific (Jackson 2016) to be generalized from a survey. This work would fill an important gap in the literature, as repetition of non-propositional units – such as punctuation and emojis – is not well understood even outside of CMC. Though the subject pool for the survey conducted here was too small for many of these questions to be answered, with a larger population and more refined focus, further perception surveys could make ground in answering questions of pragmatic functionality. Another direction not explored in the present thesis is that of humor and its function within CMC; many of

the punctuation strings observed seem to function in part as markers of humor, which could predicate their use as pragmatic markers, or be otherwise entangled. The (re)creation of suprasegmental markers of humor or for humor, such as rhythm, pauses, and pitch, could be a productive line of inquiry in elucidating the functions of nonstandard punctuation in CMC.

With that said, the present research has laid the groundwork for further work on punctuation and repetition in CMC. The investigation of pragmatics should and must extend into CMC, as a complete understanding of the creation and transmission of pragmatic meaning is not possible without it; the same underlying principles are at work, although they manifest differently. This thesis has contributed to the field's understanding of the pragmatics of CMC by examining the affective scope of punctuation. It has also provided further evidence for paralinguistic information as being *linguistic* rather than only auditory or visual. The corpus compiled here also provides an interesting overview of punctuation within a single platform and production data therein, on which much work is still to be done. This thesis took the first steps towards understanding pragmatics in CMC through the lens of punctuation and repetition thereof, and will hopefully serve as a guide and platform from which further research on such topics can be pursued.

Appendix A. Regular Expressions

!	[^\?!\.,]![^\?!\.,]
!!	[^\?!\.,]!![^\?!\.,]
!!!+	[^\?!\.,]!!!+[^\?!\.,]
?	[^\?!\.,]\?[^\?!\.,]
??	[^\?!\.,]\?\?[^\?!\.,]
???+	[^\?!\.,]\?\?\?+[^\?!\.,]
.	[^\.\?!,]\.[^\?!,\.]
..	[^\.\?!,]\.\.[^\?!,\.]
…	[^\.\?!,]\.\.\.[^\?!,\.]
….+	[^\.\?!,]\.\.\.\.+[^\?!,\.]
,	[^\.\?!,],[^\?!,\.]
,,	[^\.\?!,],,[^\?!,\.]
,,,	[^\.\?!,],,,[^\?!,\.]
,,,,+	[^\.\?!,],,,,+[^\?!,\.]
?!+|!?+	[^\?!\.,](!|\?)+(!|\?)+[^\?!\.,]
,.+|.,+	[^\?!\.,](\.,|,\.)(\.,|,\.)*[^\?!\.,]

Appendix B. Survey Stimuli

Question 1

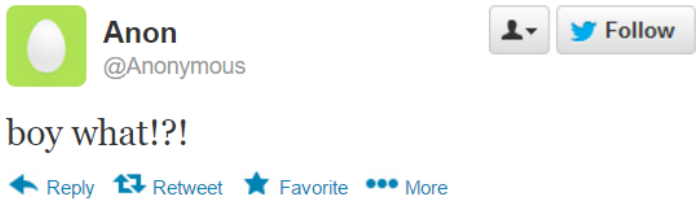
whos the boss? | whos the boss?? | whos the boss???

Question 2

how are you editing your ig stories?! | how are you editing your ig stories?!? | how are you editing your ig stories?!?!

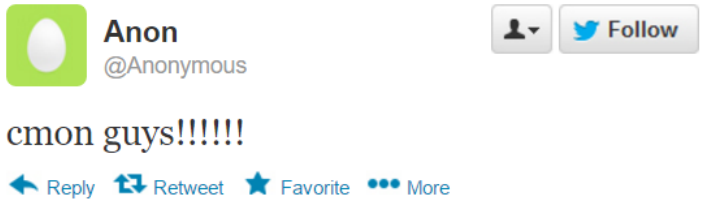
Question 3

boy what!? | boy what!?! | boy what!?!!



Question 4

cmon guys! | cmon guys!! | cmon guys!!! | cmon guys!!!!!!



Question 5

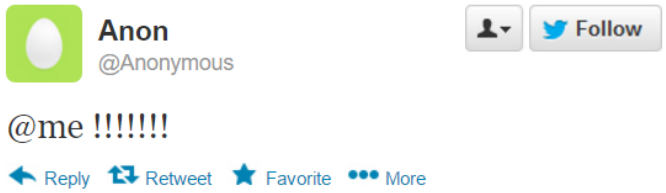
like do your job? | like do your job?? | like do your job???

Question 6

people really hate apple juice? | people really hate apple juice?? | people really hate apple juice??? | people really hate apple juice??????

Question 7

@me ! | @me !! | @me !!! | @me !!!!!!!



Question 8

that is not okay!? | that is not okay !?! | that is not okay !?!!

Question 9

I NEED THIS SHIRT! | I NEED THIS SHIRT!! | I NEED THIS SHIRT!!! | I NEED THIS SHIRT!!!!!

Question 10

where is this?! | where is this?!? | where is this?!?!

Accompanying questions

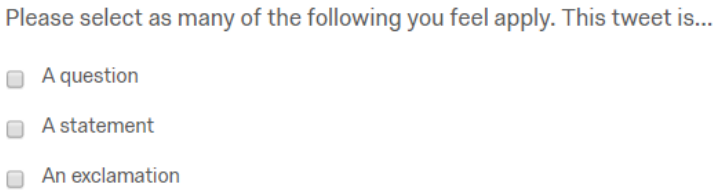


Figure B1. The first question participants were asked for each stimulus.

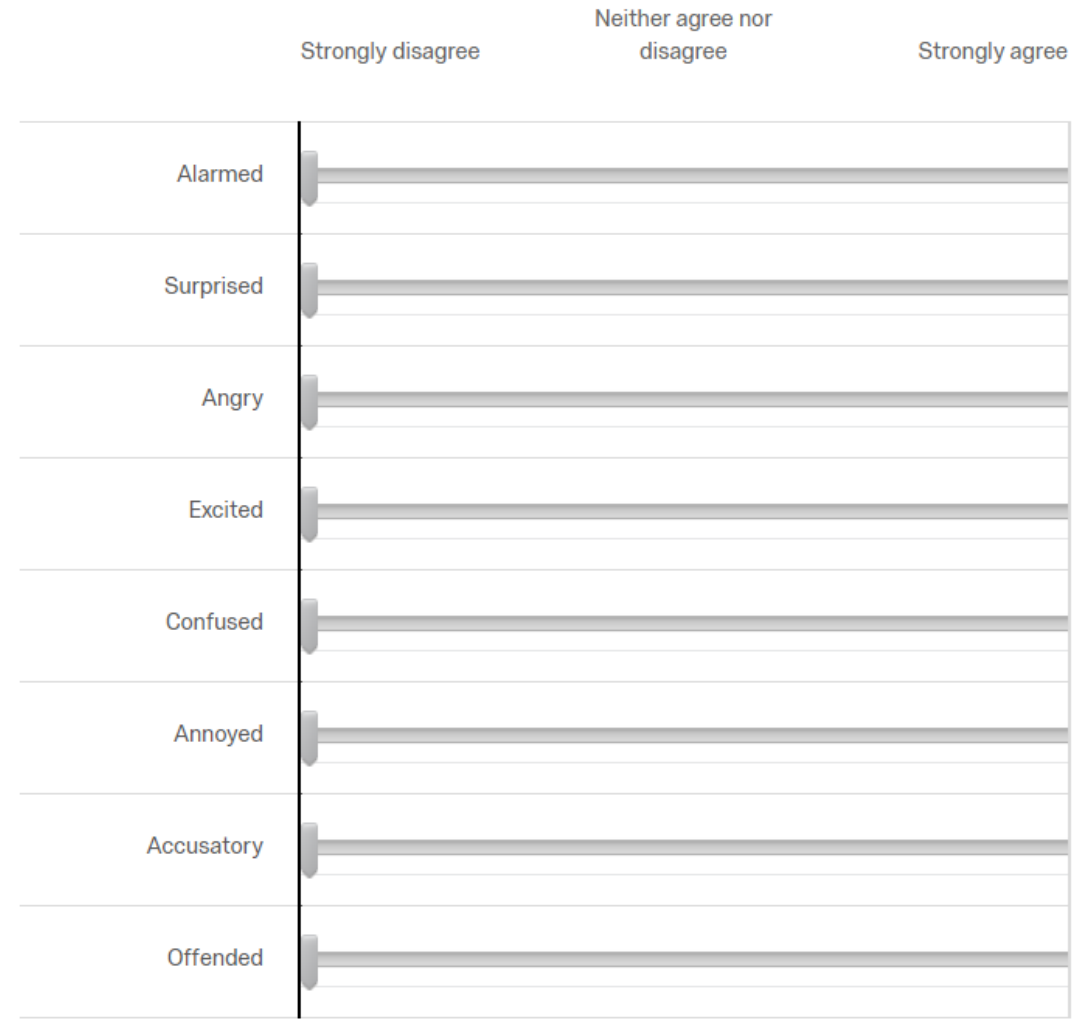


Figure B2. The sliders and emotion list as participants saw it. This came immediately after each rating of sentence type.

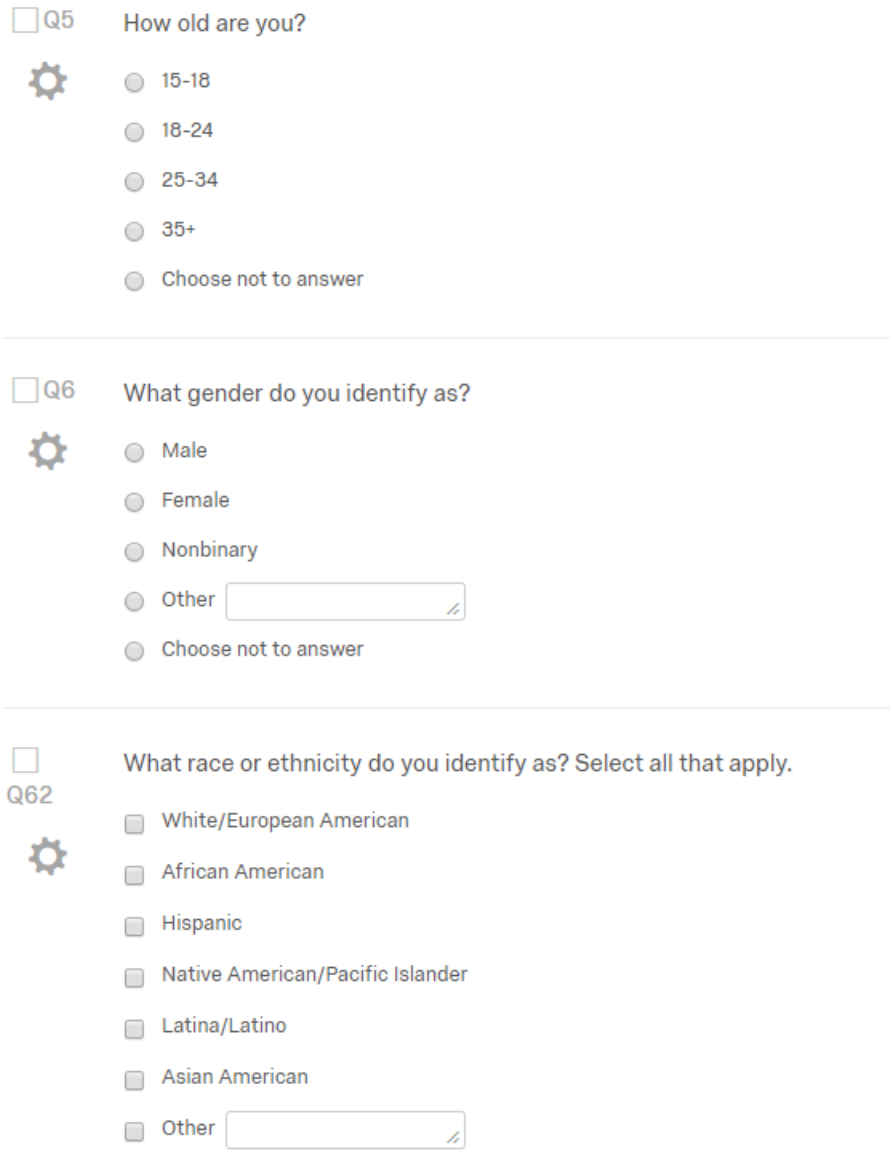


Figure B3. Demographic questions participants were asked to answer.

Appendix C. Participant Demographics

I	Platforms Used	Most Used	Social Media Time/week	Age	Gender	Race
8	Twitter, Facebook, Instagram	Instagram	2-4	18-24	Female	White/European American
9	Facebook, Instagram	Snapchat	6+	18-24	Female	White/European American
10	Twitter, Facebook, Instagram, Reddit	Reddit	6+	25-34	Male	White/European American
11	Facebook, Instagram	Facebook	0-2	25-34	Male	African American
12	Twitter, Facebook, Instagram	Facebook	2-4	35+	Female	African American
13	Facebook, Tumblr, Reddit	Facebook	4-6	25-34	Female	White/European American
14	Facebook, Tumblr, Instagram	Instagram	0-2	25-34	Female	White/European American
15	Facebook	Facebook	2-4	Choose not to answer	Female	African American
16	Facebook, Instagram	Facebook	0-2	Choose not to answer	Female	African American
17	Twitter, Facebook, Instagram, Snapchat	Twitter	4-6	18-24	Female	African American
18	Twitter, Facebook, Instagram	Facebook	2-4	25-34	Nonbinary	African American
19	Facebook, Instagram	Facebook	2-4	18-24	Male	African American
20	Facebook, Tumblr, Instagram, Other	Instagram	0-2	18-24	Male	African American
21	Twitter, Facebook, Instagram, Snapchat	Snapchat	2-4	18-24	Female	Asian American

22	Twitter, Facebook, Instagram	Twitter	0-2	18-24	Female	White/European American
23	Twitter, Instagram, Snapchat	Twitter	2-4	18-24	Male	White/European American
24	Twitter, Facebook, Reddit, Snapchat	Twitter	0-2	18-24	Male	White/European American
25	Twitter, Facebook, Instagram	Instagram	0-2	18-24	Male	White/European American
27	Facebook, Instagram, Other	Other	2-4	18-24	Female	White/European American
28	Twitter, Facebook, Instagram	Twitter	4-6	18-24	Female	Black
29	Twitter, Facebook, Tumblr, Instagram, Reddit	Reddit	6+	18-24	Male	African American
30	Twitter, Facebook, Tumblr, Reddit	Tumblr	2-4	18-24	Female	Asian American
31	Twitter, Facebook, Instagram, Reddit	Instagram	2-4	18-24	Male	White/European American

References

Aamoth, D. (2011, May 10). A brief history of skype. *TIME*. Retrieved from http://techland.time.com. Last accessed June 9, 2018.

Androutsopoulos, J. (2000) Non-standard spellings in media texts: the case of German fanzines. *Journal of Sociolinguistics* 4:4, 514-533.

Androutsopoulos, J. (2014). Languaging when contexts collapse: Audience design in social networking. *Discourse Context & Media, 4-5*, 62-73.

Anthony, L. and Hardaker, C. (2016). FireAnt (Version 1.1.4) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/. Last accessed February 1, 2018.

Baym, N. (1995). The emergence of community in computer-mediated communication. In Steve Jones ed., Cybersociety: Computer-Mediated Communication and Community. Thousand Oaks, CA: Sage, 138-63.

Bamman, D., Eisenstein, J. and Schnoebelen, T. (2014). Gender identity and lexical variation in social media. *J Sociolinguistics, 18*: 135–160. doi:10.1111/josl.12080.

Carey, J. (1980). Paralanguage in computer mediated communication. In *Proceedings of the 18th annual meeting on Association for Computational Linguistics* (ACL '80), Gary Hendrix (Ed.). Association for Computational Linguistics, Stroudsburg, PA, USA, 67-69. DOI=http://dx.doi.org/10.3115/981436.981458. Last accessed July 25, 2018.

Carlson, N. (2010, March 5). At last — the full story of how Facebook was founded. *Business Insider*. Retrieved from http://www.businessinsider.com. Last accessed July 25, 2018.

Cho, T. (2010). Linguistic Features of Electronic Mail in the Workplace: A Comparison with Memoranda. *Language@Internet, 7*, article 3. (urn:nbn:de:0009-7-27287)

Daft, R. L. and Lengel, R. H. (1984). Information richness: A new approach to managerial behavior and organization design. In B. M. Staw and L. L. Cummings ed., *Research in organizational behavior* (Vol. 6, 191-233). Greenwich, CT. JAI Press.

Darics, E. (2012). Instant Messaging in Work-based Virtual Teams: the Analysis of Non-verbal Communication Used for the Contextualisation of Transactional and relational Communicative Goals. Unpublished PhD, University of Loughborough

Darics, E. (2013). Non-verbal signaling in digital discourse: The case of letter repetition. *Discourse, Context, and Media, 2*(3), 141-148. https://doi.org/10.1016/j.dcm.2013.07.002. Last accessed July 25, 2018.

Dresner, E., & Herring, S. C. (2010). Functions of the non-verbal in CMC: Emoticons and illocutionary force. *Communication Theory, 20*, 249-268. Preprint: http://ella.slis.indiana.edu/~herring/emoticons.pdf. Last accessed July 10, 2018.

Eisenstein, J. (2015), Systematic patterning in phonologically-motivated orthographic variation. *J Sociolinguistics, 19*: 161–188. doi:10.1111/josl.12119.

Eisenstein, J., B. O'Connor, N. A. Smith, and E. P. Xing (2010). A latent variable model for geographic lexical variation. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, 1277-1287.

Eisenstein, J., B. O'Connor, N. A. Smith, and E. P. Xing (2014). Diffusion of lexical change in social media. *PLoS ONE*, 9, November 2014.

Eckert, P. (2008). Variation and the indexical field. *Journal of Sociolinguistics 12*(4), 453-476.

Gillen, J. & Merchant, G. (2013), Contact calls: Twitter as a dialogic social and linguistic practice. *Language Sciences*, 35, 47-58, http://dx.doi.org/10.1016/j.langsci.2012.04.015. Last accessed July 25, 2018.

Gunraj, D. N., Drumm-Hewitt, A., Dashow, E. M., Upadhyay, S. S. N., and Klin, C. M. (2015). Texting insincerely: The role of the period in text messaging. *Computers in Human Behavior, Volume 55*, Part B: 1067-1075. https://doi.org/10.1016/j.chb.2015.11.003. Last accessed July 25, 2018.

Hale, B. (2015, June 16). The History of Social Media: Social Networking Evolution!. *History Cooperative*. Retrieved from http://historycooperative.org. Last accessed July 9, 2018.

Hardaker, C. (2015). "I refuse to respond to this obvious troll": an overview of responses to (perceived) trolling. *Corpora, 10*(2), 201-229. DOI: 10.3366/cor.2015.0074.

Herdağdelen, A. (2013). Twitter n-gram corpus with demographic metadata. In *Language Resources & Evaluation, 47*: 1127-1147. https://doi.org/10.1007/s10579-013-9227-2. Last accessed July 9, 2018.

Herring, S. C. (2001). Computer-mediated discourse. In D. Schiffrin, D. Tannen, & H. Hamilton (Eds.), *The Handbook of Discourse Analysis* (pp. 612-634). Oxford: Blackwell Publishers. http://ella.slis.indiana.edu/~herring/cmd.pdf. Last accessed July 25, 2018.

Herring, S. C. (2007). A faceted classification scheme for computer-mediated discourse. *Language@Internet*. http://www.languageatinternet.org/articles/2007/761.

Herring, S. C., & Androutsopoulos, J. (2015). Computer-mediated discourse 2.0. In D. Tannen, H. E. Hamilton, & D. Schiffrin (Eds.), *The handbook of discourse analysis*, Second Edition (pp. 127-151). Chichester, UK: John Wiley & Sons.

Herring, S. C., Stein, D., & Virtanen, T., Eds. (2013). Introduction to the pragmatics of computer-mediated communication. *Handbook of pragmatics of computer-mediated communication* (pp. 3-31). Berlin: Mouton.

Iorio, J. (2009). Effects of audience on orthographic variation. *Studies in the Linguistic Sciences: Illinois Working Papers*, vol. 2009, 127-140.

Jackson, R. C. (2016). The pragmatics of repetition, emphasis, and intensification (Doctoral Dissertation).

Kalman, Y. and D. Gergle. (2014), Letter repetitions in computer-mediated communication: A unique link between spoken and online language. *Computers in Human Behavior, 34*: 187-193. http://dx.doi.org/10.1016/j.chb.2014.01.047. Last accessed July 9, 2018.

"Myspace." *Crunchbase* (n.d.). Retrieved from https://www.crunchbase.com/organization/myspace#section-overview. Last accessed July 20, 2018.

Lave, J. & Wenger, E. (1991). *Situated Learning: Legitimate Peripheral Participation.* Cambridge: Cambridge University Press.

Lin, Y. (2016). Non-standard capitalisation and vocal spelling in intercultural computer-mediated communication. *Corpora, 11*(1), 63-82. doi:10.3366/cor.2016.0085.

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

Pavalanathan U. and Eisenstein, j., (2015). Audience-modulated variation in online social media. *American Speech 90*(2): 187-213; doi:10.1215/00031283-3130324.

Petronzio, M. (2012, October 25). A Brief History of Instant Messaging. *Mashable*. Retrieved from https://mashable.com. Last accessed July 9, 2018.

Raclaw, J. (2006). Punctuation as Social Action: The Ellipsis as a Discourse Marker in Computer-Mediated Communication. *Annual Meeting of the Berkeley Linguistics Society, 32*(1), 299-306. doi:http://dx.doi.org/10.3765/bls.v32i1.3469. Last accessed July 25, 2018.

Rosen, A. (2017, November 7). Tweeting Made Easier. *Twitter Blog*. Retrieved from https://blog.twitter.com. Last accessed July 9, 2018.

Rosen, A. and Ihara I. (2017, September 26). Giving you more characters to express yourself. *Twitter Blog*. Retrieved from https://blog.twitter.com. Last accessed July 9, 2018.

RStudio Team (2016). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL http://www.rstudio.com/. Last accessed July 20, 2018.

Schandorf, M. (2012). Mediated gesture: Paralinguistic communication and phatic text. *Convergence: The International Journal of Research into New Media Technologies, 19*(3) 319-344.

Sinclair, J. (1991). *Corpus concordance collocation*. (Oxford: Oxford University Press).

Smith, A., and M. Anderson (2018). Social media use in 2018. In *Pew Research Center*.

Squires, L. (2010). Enregistering internet language. *Language in Society 39*, 457–492. DOI:10.1017/S0047404510000412.

Squires, L. (2012a). Whos punctuating what? Sociolinguistic variation in instant messaging. In Alexandra Jaffe, Jannis Androutsopoulos, Mark Sebba, & Sally Johnson (Eds.), *Orthography as Social Action: Scripts, Spelling, Identity and Power*, 289-324. DeGruyter (Language and Social Processes).

Squires, L. (2012b). "Twitter: Design, discourse, and the implications of public text." The Routledge handbook of language and digital communication. Ed. Alexandra Georgakopoulou and Tereza Spilioti. London: Routledge Handbooks Online, 2015. 239-55. Print.

Squires, L. 2016a. Computer-mediated communication and the English writing system. In Vivian Cook and Des Ryan (Eds.), Routledge Handbook of the English Writing System, 471-486. Routledge.

Squires, L. (Ed.). 2016b. English in Computer-Mediated Communication: Variation, Representation, and Change. *De Gruyter (*Topics in English Linguistics 93).

Stenberg, D. (2011, March 29). History of IRC (Internet Relay Chat). Retrieved from https://daniel.haxx.se/irchistory.html. Last accessed July 9, 2018.

Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work* (Amsterdam/Philadelphia: Benjamins).

Twitter Tweet Generator. (n.d.). Retrieved January 11th, 2018 from http://simitator.com/generator/twitter/tweet. Last accessed July 9, 2018.

Vandergriff, I. (2013). Emotive communication online: A contextual analysis of computer-mediated communication (CMC) cues. *Journal of Pragmatics 51*, 1-12. https://doi.org/10.1016/j.pragma.2013.02.008. Last accessed July 25, 2018.

Walther, J. B. (1992). Interpersonal effects in computer-mediated interaction: A relational perspective. *Communication Research 19*, 52-90. DOI: 10.1177/009365092019001003.

York, A. (2017, March 6). *Social media demographics to inform a better segmentation strategy*. Retrieved on April 4th, 2018, from https://sproutsocial.com/insights/new-social-media-demographics/#twitter. Last accessed July 9, 2018.

VITA

Bachelor of Arts; English, North Carolina State University

Master of Arts (Expected); Linguistic Theory and Typology, University of Kentucky

Elizabeth M. Wright