SHE, HE AND THEY TRENDING ON TWITTER: POLYVOCAL PRONOUNS AND MORE-PUBLIC MESSAGES

A thesis
submitted to the Faculty of the
Graduate School of Arts and Sciences
of Georgetown University
in partial fulfillment of the requirements for the
degree of
Master of Arts
in Linguistics

By

Elizabeth M Merkhofer, B.A.

Washington, DC
April 20, 2013

SHE, HE AND THEY TRENDING ON TWITTER:
POLYVOCAL PRONOUNS AND MORE-PUBLIC MESSAGES

Elizabeth M Merkhofer, B.A.
Thesis advisor: Anna M Trester, Ph.D.

ABSTRACT

This paper uses ethnographic and quantitative methods to examine the use of standard, third-person personal pronouns and 'singular they' for a specific referent in a Twitter corpus. The corpus, collected from the Twitter Search API, includes 9031 Tweets from the then-trending hashtag '#oomf' [one of my followers]; the 1139 of these found to contain singular masculine-, feminine- or 'singular they'-type third-person pronouns linked to #oomf are studied here. Previous studies find generic-related features of the referent to differentiate 'singular they'. These features, however, do not account for pronoun variation among the tokens in these data, only some of which include additional gendering of referents and which are consistently of a referent type considered by previous research to be unlikely with 'singular they'. Rather, a pattern emerges at the corpus level: two texts including 'singular they' are disseminated by many users, while Tweets containing standard pronouns accumulate only small-scale social action. It is suggested that through users' social actions, 'singular they'-including Tweets become polyvocal and comparatively more public. The association of 'singular they' with low individuation and prototypicality emerge here at the level of discourse and social behavior, rather than predicting the use of a certain pronoun in sentences.

ACKNOWLEDGEMENT

I have many people to thank for their many ways of supporting the research and writing of this thesis. First is Anna Marie Trester, my advisor in this process, who introduced me to ethnography and talked me through many more inspirations and analyses than have made it to this final version. Second, I would like to recognize how interested and willing to engage the entire Georgetown Linguistics department has been, and I thank my peers especially for supporting me through the many singular they's they Tweeted, emailed, and Facebooked and recounted to me. My classmate and friend Dan Simonson deserves special recognition for getting me off the ground with Python; without him, I'd still be trying to run that if-loop without a colon. Finally, I am grateful for the support of my family and friends who would probably just as soon never hear another singular they. I owe so much to my parents for valuing education so highly.

TABLE OF CONTENTS

I. Introduction

Who is #oomf?

- #oomf is a cutie.
- I have NO respect for #oomf.
- #Oomf be makin ugly faces ,
- Wonder wat going on in #oomf life

#oomf is a Twitter hashtag - pronounced /umf/ or perhaps O-O-M-F - that is used to talk (or Tweet) about a specific person. Literally, it abbreviates "one of my followers", but in practice, it has acquired a more specialized, somewhat indirect meaning – according to my interpretation, something of an online rehash of the offline phrase "a certain someone." From the hashtag, which seems to collect a steady stream of Tweets but periodically spike in popularity, a wealth of portrayals of users' friends and enemies, objects of lust and loathing emerge.

In this thesis, I explore a collection of Tweets about #oomf to see how "singular they" is used to talk about specific people. Drawing from qualitative, ethnographic consideration of Twitter discourse and quantitative approaches to examine those Tweets, I examine the third person pronouns linked to #oomf in about 30 minutes of Tweets. The findings of previous studies of THEY, which mostly emerge from assumptions that THEY is used to fill a "gap" in the English pronominal system for non-sexist generics, are found to apply poorly to the data, once variation cannot be correlated with the common antecedent and type of referent. Similarities, rather than differences, are found among tokens. Furthermore, measuring the co-occurrence of further linguistic and textual features with pronouns, a common practice in Twitter studies, seems too naïve for a corpus rife with shared texts. Rather, it is the discovery of the intertextual

connections among Tweets that offers clues to the distinction of singular ‘they’ from the standard third-person pronouns. The large scale of dissemination of two particular texts with “they” linked to #oomf suggest that THEY comes to be used for prototypicality or low individuation occur at the level of the corpus, and that these features, indistinguishable in individual sentences, emerge as the text is made public through the social action of many users. That is, if messages with THEY are unique from those with SHE or HE because their referents are used to illustrate a type or to stand in for many individuals without particularly distinguishing them, it appears to come to do so through large-scale use rather than facts traceable to individual sentences.

The following proceeds by introducing the Twitter platform, reviewing previous literature, introducing the dataset, offering results discovered at both the Tweet level and the level of related Tweets, and finally, offering conclusions and directions for further study.

II. What is Twitter?

Twitter is an extremely popular social media platform. The platform was founded in 2006, and reached 500 million user accounts in June of 2012, 140 million of those in the United States (Semiocast 2012). Users access Twitter through a web interface at Twitter.com, through mobile applications on tablets or smartphones, or via SMS. The platform describes itself, on its website, as a real-time information network; it is also popularly described as a "micro-blogging" service, micro in the sense that users' contributions are very brief, and blog-like in the expectation that users Tweet about themselves. In order to introduce the platform and the social organization of my corpus to the reader, I first explain what it means to be a user, then describe the

publication and republication of texts, giving a partial account of how those texts appear throughout the platform, focusing in on organization through hashtags.

Though many pages on Twitter can be viewed without any authorization, active participation in the platform may only be done through logged-in accounts. Sign-up is free and open to anyone able to access the Twitter.com site. The process is very quick and requires a minimum of personal information. New users are prompted to enter a real name (my experience with the site indicates that these often do not follow a 'first name, last name' or 'company name' format, but may be used for self-description or parody like many other social media handles). Users also enter an email address, with which they confirm the creation of the new account, and a password. Finally, they must create a username with any combination of letters and numbers prefaced by '@'. This username, unlike the "real name," must be unique in the platform. These fields are sufficient to create a new identity in the platform, but the user may offer more information for the profile that is created - for example, upload a small avatar, enter a biography of up to 160 characters, indicate location. From this profile, the user may profile may then asymmetrically form connections to others or be connected with: any user can "follow" any other user, meaning receive the Tweets of that user, and be "followed" by other users, meaning broadcast their Tweets to that user. These identities may also be used actively within the platform to create or propagate texts.

A Tweet is the primary mode of interaction in Twitter, and any user may Tweet. A Tweet begins with a user's purely textual contribution of up to 140 Unicode characters. When that Tweet is submitted, it becomes integrated into the Twitter platform. For one, parts of the user-entered text become interactive, creating links between the Tweet and other texts, users, and

sites. URLs in the text become hyperlinks. Alphanumeric strings prefaced with @ become hyperlinks to the profile page of that @username, whether occupied or not; this is called an @-mention. Alphanumeric strings prefaced with # become hashtags, which, as will be explained below, are a system of Tweet organization within the platform. This research is built around texts brought together by a hashtag, which in addition to unifying them in topic, connects the texts to each other through this interactivity. As a whole, the text becomes integrated in a standardized and interactive display format used in the official website and app and required of any third-party client (Twitter 2012a). Each Tweet is encapsulated in a rectangular space, with the user's avatar to the extreme left, and the user's "real name" and @username above it, all of which link back to the user's profile. A timestamp to the extreme right links to a permanent URL for that Tweet. The user Tweet text is displayed in the middle of the box. A row of "Tweet actions" can be seen along the bottom (on the website, when a user "mouses over" the Tweet). This is illustrated in Figure 1 below, a Tweet by the popular musician Lady Gaga (real name) or @ladygaga (@username) on January 31. The menu of Tweet actions for this Tweet is expanded by a mouse over, and the options (reply, retweet etc.) are available.

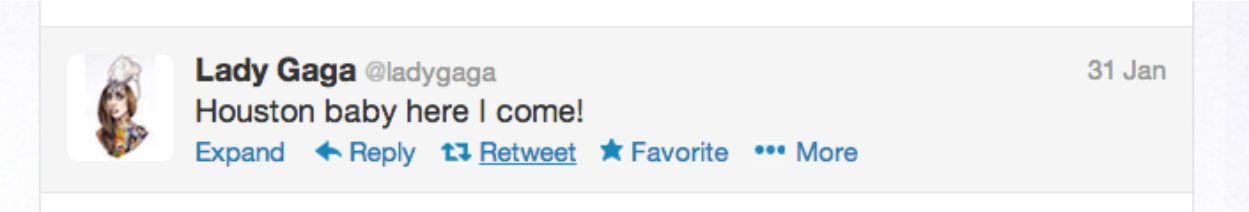

Figure 1. This screenshot of a Tweet on Twitter.com shows its elements and available interactions.

Not all Tweets, however, are unique, original texts; rather, many are reproduced from other users, especially through the practice of retweeting. The Retweet action re-publishes a Tweet under a new name, and it accounts for over half of all Tweets in the corpus analyzed in

this thesis (as is true of any hashtag corpus I have examined thus far). A variety of conventions have emerged for marking texts as copied Tweets and, optionally, providing attribution, such as typing "RT @[USERNAME]" or "via @[USERNAME]:" before the copied text (see boyd, Lotan and Golder 2010). These Tweets are then displayed just the same as any other original Tweet, including a hyperlink to the original user's page through the @-mention in the typed Retweet tag, and any other hashtags or links also active in the copy. These conventions were supplemented in 2009 when the platform added a native Retweet function, which mobile apps have gradually come to support. These Retweets display slightly differently in some user interfaces, such as Twitter.com and its official app: the original Tweet, including the username, avatar, and real name of the original author, is reproduced, with a line underneath the Tweet text that identifies it as a Retweet and the Retweeting user: "Retweeted by [REAL NAME]." Some other apps display the avatar of the Retweeting user instead of the original author, and some display “RT @[USERNAME]” before the whole Tweet. At times, users reproduce Tweets by copying and pasting the text without providing any link to the previous text.

These Tweets and Retweets are displayed across the site in several collocations, all of which are updated in real time. On each page, Tweets are displayed in reverse chronological order with the newest Tweets at the top, and a notification bar appearing above all Tweets that will allow the user to load new Tweets as they are produced. Some streams are organized by user identity. The profile page of any user displays the stream of Tweets and Retweets by that user, with the newest at the top. Each user has a "Home" page, where the Tweets of all the users they follow are displayed in reverse chronological order (alternate views, organized by

conversations created with the Reply action and @-mentions, are available for some Tweets through extra clicks).

This corpus is collected based on Tweets that would be included in hashtag streams, which bring together topically-tagged Tweets. In these streams, viewable as real-time search pages on Twitter, users exploit the hashtag as an organization system for Tweets. Tweets including the same hashtag are brought together by the platform on a search page stream for that hashtag. The searchability of the hashtag (whereby every iteration of a # followed by any alphanumeric string is hyperlinked to its search page) was added by the platform in 2009. However, the convention began before that, as users prefaced words with "#" in order to attach topical metadata to a Tweet, simultaneously creating and indexing content (for example, "#fail" at the end of a Tweet about a comically unsuccessful attempt to do something). A hashtag search page may display few or no Tweets if the particular hashtag string has been used only idiosyncratically or has not been used within the past two weeks. However, some hashtags develop enormous popularity and become the site of active interaction. They bring together and prompt interactions between users and texts.

The participants in a hashtag are distinct from the Twitter community in general. Participation in a hashtag is self-selecting, in that users participate only in hashtags they find personally relevant or that spark a personal reaction, and affiliative, in that a hashtag is only one of many ways to express a topic but is exploited by multiple users to bring their contributions together. The populations Tweeting about a hashtag cannot be taken to represent those using Twitter in general. However, in the case of #oomf, studied here, this offers unique possibilities of study. In what follows, I'm able to use the texts that emerge

around a hashtag not only to focus on one linguistic form, but also to capture texts in interaction with each other. The affiliative use of #oomf redefines it as an antecedent to introduce a specific referent, allowing the environment of Tweets and their antecedent to be held constant. Furthermore, the texts that emerge from this affiliation sometimes make apparent deeper affiliation with each other, allowing me to trace linguistic form in use between many users.

III. Literature Review

This literature review proceeds by first surveying what has been written about singular they, which as of yet has mostly emerged as a reaction against sexist generics (i.e. “generic he”). Then, it reviews trends in studies of online communication that demand that social media data be examined based on the actions and identities present in platforms, rather than offline categories. Finally, it outlines an approach to words through context.

i. Singular they

This paper draws from previous research on singular they, which often focuses on gender in pronouns. This previous research often focuses on exposing gender persistent in so-called epicene (including no indication of gender), gender neutral, or sex-inclusive pronouns, such as the once-prescribed use of “he” in a sentence like the following where the gender of the referent is actually not yet determined: “If someone calls, tell him I just left.” Furthermore, the original research performed in this thesis asks about the relevance of social relations other than gender in the differential use of third person singular pronouns, and it would not like to presume that “singular they” is itself gender neutral. As such, calling third-person singular

pronouns “masculine”, “feminine” or “epicene” seems to beg the question. The term “singular they” is inelegant. And so I begin this paper by establishing the convention of writing the third-person singular pronouns that are the object of my study in the nominative case IN CAPS to signify the entire class of pronouns. That is:

	shall stand for
SHE	she / her /hers / herself, etc.
HE	he / his / him / himself / hisself, etc.
THEY	they / them / their / themselves / themself, etc

THEY has been the object of much study since at least the 1970’s, when the issue of sexism in generic pronouns (specifically, HE in this sense) became a topic of feminist, prescriptivist, and thereafter, descriptivist linguistic inquiry. These early feminist writings, in their critique specifically of generics, find sexism deeply embedded in language and the processes of prescription. They commonly observe that the history of THEY traces back to generics in such respected old tomes as Shakespeare’s ‘Comedy of Errors’, long before its proscription in favor of a more “Latin-like” HE in the 1800’s. They write that the use of a masculine pronoun can never be truly neutral, that HE necessarily renders women "invisible and silent” (Baron 1986, 100). They find the distinction between masculine and feminine gender in language to be imposed, rather than necessary, and find that it reflects and reproduces societal hierarchies (see for example Cameron 1992: 88) Cognitive studies corroborate the point, finding that HE used generically and for explicitly sex-inclusive groups nevertheless brings males to mind either exclusively or before females (see, for example, Moulton et al 1978). The point that gender-specific pronouns cannot stand in for referents whose gender does not match is illustrated especially clearly as McConnell-Ginet find the

"generic she" to be acceptable only when feminine referents have special salience (2011b, 196). The studies, however, tend to share the premise that THEY is motivated by lack of information about the gender of the referent, a condition that likely does not apply to most texts in the corpus studied here.

Much ink has been spilled as authors theorize a gap in the English pronominal system. Many authors write of a need for a third-person pronoun that does not specify gender, conceived mostly for generic purposes. In one functional linguistic account, Weidmann (1984), upon illustrating several sentences that do not determine gender but which require a third-person, singular (and emphatically, not plural) pronoun puts it as follows: "Two principal ways of stopping our gap are open: either we invent a new pronoun to express the combination of features described above, or we make do with the existing pronouns" (60).

Prescriptive solutions that work toward gender neutrality have had little success in the English language. Pauwels (2003) observes that attempts to reform sexism in the English language have turned almost without debate to strategies of gender-neutralization (in comparison to e.g. German, which has relied more heavily on feminization), which have met with limited success (559). Various neologisms suggested (most since the 1970's) to achieve sex-neutrality for singular, epicene or sex-inclusive pronouns are written about by Baron as "the Word that Failed" (1986, 190-216). Gender neutral noun phrases, where newly coined, are hardly stable. Where actor and actress have been leveled to actor, "woman actor" re-emerges in use (Harré and Mühlhäusler 1990, 240); newly-minted 'gender neutral' forms (e.g. "chairperson") come, in practice, to be used as the marked term mostly or exclusively for females while men continue to be referred to with the original form of the word ("chairman")

(Ehrlich and King 1994). In literature, several authors have attempted to create stories about characters without Gender, but as Livia observes in her 2001 survey of several such English-language novels, the characters produced are “disjointed” and “distant.”

These failures of sex neutrality are very much in contrast to the persistent and common use of THEY; here, rather than finding speakers “making do” with THEY, THEY is found to occur frequently and even to seem to be the optimal choice for some speakers and referents. Though the question and gap remain open from a prescriptivist lens, considerations of language in use show a different side of the story. Miller and Swift suggested in 1976 that THEY was most likely to succeed in the proposed “gap,” because it was "already commonly used both in speech and writing” with both indefinite pronouns and lexical noun phrase referents of "indeterminate or inclusive gender" (135). Baron found the pronoun met the least resistance when used in places where it is syntactically singular but semantically plural, like with indefinite nouns like "person, someone, or everyone" (1986, 193-6). Stringer and Hopper's study of spoken English, based on data spanning from the 1960's to the 1990's, finds such infrequent use of generic HE that the authors question whether the form was ever "an unmarked usage in English conversational interaction" (1998, 209). Cameron observed in 1992 that THEY was used "in some spoken contexts almost invariably" and "generic he" not at all (95). Further work points to the lexical availability of the pronoun (for example McKay 1980’s analysis of the suitability of the pronoun for prescriptivist purposes). In studies of college compositions including generic, third person pronouns, Myers 1990 finds that THEY is used consistently by the largest percentage of students and does not seem to represent an error in number. Holmes (1998) finds that 80% of the non-specific referents in a corpus of (mostly informal) New Zealand speech are

pronominalized with THEY and dismisses HE as so infrequent as to be a “pseudo-generic” (30). These studies, together, suggest that THEY is very much in use as a third-person pronoun, but, emerging from the premise that THEY is generic, do not attempt to locate it in broader contexts.

Since THEY is so often studied in contrast with generic HE, the research often produces conclusions about gender, showing that THEY use is affected both by the gender of the speaker and by gender stereotypes about the referent. Together, these studies show that women tend to use THEY more, perhaps to talk about generic men or perhaps to talk euphemistically about women, and in cognitive tests, is interpreted as being about men. Multiple studies of generics find that HE and SHE are overwhelmingly used when information about gender is available (for example, for sex-exclusive groups (Newman 1997, Balhorn 2001), or sex-stereotyped NP antecedents (Matossian 1997, Balhorn 2001)). When THEY is used, it is disproportionately for masculine referents. Matossian (1997) illustrates a “male bias” in men’s use of THEY IN elicitation tasks where subjects completed the missing generics in written sentences. Though the gender roles stereotypically associated with the gender of a given antecedent are typically matched by the gendered pronoun, a closer inspection of uses of THEY shows that the uses are more frequently for stereotypically masculine referents and less frequent for stereotypically feminine referents. Female respondents did not show this pattern. This tendency of THEY toward masculine referents is also reported in a 1989 study in which participants drew and talked about sentences containing generic pronouns. Each generic pronoun was interpreted with male figures more than with female figures, and with genderless figures the least. THEY produced even more masculine representations and fewer sex-neutral representations than sentences with “he and she”, except in female participants who themselves used non-sexist

pronouns (Khosroshahi 1989). These studies suggest a reevaluation of introspective studies that analyze THEY to lack gender marking; though formally it may not, authors seem to consistently find referent gender relevant to its use and its interpretation. Several further studies further support the importance of speaker gender. Pauwels finds that women use more "non-sexist" alternatives including THEY and almost no HE (see Pauwels 2003, 564, for the synthesis of several of her studies). Myers (1990) finds that college-age women were significantly less likely to use generic HE in an elicitation task, and although their strategies to avoid it were quite mixed, she found many consistently using THEY. Balhorn (2001) reports that female authors in his newspaper corpus use a larger proportion of THEY and less generic HE, and suggests this reflects female authors' feelings about inclusion. These findings contribute to THEY research by showing that THEY shows more nuanced, gender-based tendencies than those implied by its apparent formal gender neutrality; however, they all tend to arise from studies of formally gender-neutral (if sex-stereotyped), generic contexts and the use of THEY where referent gender may be known is not yet considered as a possibility.

These and other studies have also found other, generic-related factors related to the referent, sometimes tied to the plurality of THEY, to be important - even *more* important - in predicting the use of THEY. Further, generic-related features of the discourse also condition THEY, like notional plurality and concreteness of the referent, and personalization. Newman, analyzing pronouns co-referent with singular epicene antecedents in TV interviews, correlates pronouns with further semantic features of the referent. THEY is found to be more common than HE, while other strategies like SHE and "he or she" are extremely rare. Stemming from the connection of "they" as a plural pronoun and "singular they", he introduces the concept of

“notional plurality” in his analysis of the syntactically singular antecedents and quantifiers linked to THEY at three levels of sentential “notional number,” plural (ex. “every”), singular and ambiguous (or neutral). However, upon closer analysis of the co-referent antecedents, Newman finds almost no notionally singular antecedents used with THEY and concludes that plurality is even more closely associated with THEY than gender neutrality, and it is conditioned by less concrete antecedents (470). However, when Baranowski (2002) revisits the notional number in epicene referents, while she finds THEY to be used nearly invariably for notionally plural referents, she also finds that THEY is still not only possible but the most common pronoun for generics in the other two groups. Balhorn's work with a newspaper corpus corroborates the finding that THEY is used with less concrete individuals, finding that existential antecedents (e.g. anyone/-body) are most often used with THEY, and the more notionally singular lexical NP antecedents (“a person”, “a CEO”) are most often used with HE, SHE, or "he or she”. McConnell-Ginet’s paper arguing that the use of pronouns is attached to conceptualizations of “people” and “prototypes” suggests that THEY prevents personalization because of its gender neutrality: “The reason singular *they* is not very satisfactory with definite generics and intolerable with proper names is that both personalize their referents and give them a particular identity, endow them with personality. Personal identity, personality, is in our culture closely tied to sexual identity” (2011b, 200). This echoes an intuition that characterizes McKay’s 1980 exploration of prescriptivist possibilities of THEY, as the author is reluctant to recommend the pronoun in part because the authors find it depersonalizes (generic) referents.

Finally, a few studies reach the edge of the feminist and generic underpinnings in THEY literature, uncovering uses of THEY that are not neutral or not generic. These accounts suffer

from less systematic documentation, as they seem to mostly emerge from aberrant tokens noted in corpora or from introspection. In a 2003 paper, McConnell-Ginet illustrates what she characterizes as a restricted but increasing use of THEY: “A friend of Kim's got their parents to buy them a Miata.” However, she qualifies this use, adding that THEY “is still unlikely to be used for a specific individual in many circumstances: if, for example, both interlocutors are likely to have attributed (the same) sex to that individual" (reprinted as McConnell-Ginet 2011a, 230); this seems strangely restricted, in view of subsequent research. In their corpora, Balhorn (2009) and Newman (1992, 1997) each observe the use of THEY for individuals of known gender to the speaker but with "low individuation," who appears in discourse "merely as a type:" Balhorn provides the example of a restaurant review that includes an anecdote about a patron at another table as THEY. In 2004, Balhorn writes more decisively that THEY may be substituted in discourse where HE or SHE would make the feature of gender too salient, for example, where “Somebody called while you were out and he said he’d call back later” does not express an appropriate degree of disinterest in the caller (84). Lagunoff (1997) offers a similar analysis, suggesting that THEY may be pragmatically motivated as speakers attempt to avoid providing information of poor quality.

Based on a survey of old English texts, Balhorn (2004) creates an argument that he explicitly disconnects from the feminist critique of sexism in generics, claiming instead that the syntax of English internal pressures that slowly allowed THEY to “rise” to its current place in language as an “unmarked” pronoun that allows speakers to highlight animacy of the referent while not foregrounding gender. (Unfortunately, he provides little modern documentation of this happening in non-generic contexts that would add to this discussion.) The Language Log

blog documents further examples for indefinite or unknown definite antecedents of known sex like ‘a users of a men’s restroom’ (see Zwicky 2010) and even with a personal name in a letter soliciting information about a job applicant (see Pullum 2010’s “Singular They with personal name antecedent” including the user comments judging grammaticality and suggesting that the usage is politically hypercorrect or an unedited form letter). Finally, Lagunoff, whose 1997 doctoral dissertation focuses heavily on the different types of antecedents (ex: quantifiers, definite noun phrases), opens the door to possible combinations. She comes to conclude that “antecedents of Singular they can be of any kind, including where the gender is overt or implied, except names” and with pointing (xiii). In what follows, I systematize the search for such specific referents. In order to account for the data examined here, I recognize it also as inextricably tied to the online context it emerges from.

Qualitative work recognizes that the variability of terms creates the potential for rich, social meaning. Terms of reference, especially, are encoded with social significance: Schiffrin (2006), working with multiple references over extended narratives, demonstrates how referring to a person in one way always represents choosing that term over other, available options and thereby constructs a certain type of relationship between speaker and referent. These terms of reference uses both "denotation," and subjectively colored "connotation" to construct the person talked about (46). Pronouns, too, have been shown to project social categorizations. In their 1990 book *Pronouns and People*, Mühlhäusler and Harré explore how this small unit of language creates concepts of people; they find that down to the level of grammar, “person-indicating expressions in most languages include reference to specific social relations”(5) - especially gender, but also hierarchy. This encoding of relations into pronouns is inherently

social, contingent upon practical knowledge of the world and society (53). Joining other researchers, I look to situate this in online communication.

ii. Online and Twitter communication

As online texts are increasingly the object of linguistic study, researchers have come to recognize the situated nature of such texts and the uniqueness of the social relations they represent. Rather than unproblematically extending traditional objects of analysis to internet data, attention is turned to the unique possibilities and practices of interaction from which they emerge. The following reviews sociolinguistic perspectives on gender in Internet communication and on alternatives to study of identity that have emerged in internet study. Then, it overviews writing specific to Twitter and to the interactive features of the text that will later be analyzed in this original study.

Internet researchers have clearly established that communities exist online, but also that they take unexpected or novel forms. For example, as Kozinets (2010) reports from the field of Internet ethnography, internet researchers identify indistinct boundaries to community membership in online participation; yet, users create cohesive social orders and develop norms specific to their online environments. Baym and boyd (2011) recognize online communities as complex, multi-layered sites of interaction, where interaction may be more or less public based on audience and participation that emerges; as Marwick and boyd wrote of Twitter in 2011, this continuum of contexts is even collapsed into individual Tweets, which are shown indiscriminately to followers of more or less familiarity with Twitter users in the many domains they Tweet about. In a 2011 paper, Gruzd et al wrote about the experience of Twitter through the eyes and connections of one user, concluding that the platform offered an “imagined

community" that shared norms, which was characterized by being both collective (in that Tweets are able to be accessed by the whole platform at any time) and by being personal (because users are situated and findable within their own unique identities). Wu et al (2011) introduce the term "masspersonal communication" in the context of Twitter, where communication is neither clearly mass (one to many) nor personal (one to one) but must serve both communicative functions.

Research on online communication increasingly asks how its saliencies, shaped by the unique context of platform, differ from offline communication. For example, Androutsopoulos' discourse analytic work approaches MySpace users' identity through code variation with attention to the platform's specific, interactive functions, rather than by researching users' offline macro-demographics; as he urges, analysis of language in social media must be based in an understanding of the functions and salient features of the platforms where the data is produced (2010). In the case of Twitter, the research looks to the performance of multiple interpersonal involvements in a shared environment. Specifically to Twitter, Gillen and Merchant's 2011 autoethnography arrives at the conclusion that participation in the platform implies the possibility of participation in social connections (ex: being retweeted or replied to), and a limiting of potential actions to those supported by Twitter; this openness to interaction and potential polyvocality within the whole group leads to a "partial intersubjectivity." Java et al (2009) provide an account of Twitter by asking what user's intentions are; they posit that users participate in multiple different communities split between different interests, and that they are led primarily by three intentions: information sharing, information receiving, and friend-wise communication. Marwick and boyd (2011) write about this as "context collapse" -

users negotiate person-to-person and topic-specific connections within an environment that does not allow separation of audiences and contexts; Twitter users must imagine their audiences, as they cannot know which follows read which Tweets, the often negotiate audience design based on Tweet content.

A study that contrasts THEY with gendered pronouns must ask if gender is performed differently or more easily left unspecified online; according to previous research, online communication has not made gender irrelevant but rather reinforces it in unique ways. As Herring (2003) writes, online communication creates the potential for genderless self-portrayal. Since it is text-based, online communication lacks the visual and auditory gender cues of other forms of communication. Theoretically, gender could remain anonymous or irrelevant. Still, Herring's work finds that actual practices continue to be gendered, as users make their gender overt or "give themselves away" with styles typically associated with gender, and that the gendered discourse produced exhibited the same asymmetries as offline communication. Other research suggests that the possibilities of genderless representation online actually promote greater foregrounding of gender. In Sundén's 2002 research on a text-based online program of Multi-User Dungeons, she writes that "textual talk" constitutes gender and online bodies for users. Despite a plurality of gender options, most characters conform to a male/female gender binary. Players who select other options (such as "neutral" or "plural") are characterized by repeated attention to gender and sex in their texts, and one player who prefers the pronoun "it" is referred to with feminine pronouns by a familiar user. Sundén writes that the environment of uncertainty about physical bodies leads users "to textually re-inscribe familiar categories on the level of sex and gender, to insist on a system of recognizable differences. (301)" In 2012,

Bamman et al locate gender in Twitter users' patterns of lexical items. Though some features do not follow sex-based patterns precisely as found in offline communication, most users' gender can be accurately predicted from an analysis of their style. Interestingly, those whose cannot tend to have Twitter connections mostly to members of the opposite sex, whose style they mirror. This suggests that Twitter, which does not formally solicit gender information from users, is not a genderless environment, but rather emergent norms in the community lead to self-revelation.

Much more, research on Twitter communication has focused on creation of coherence through the textually based tagging system that connects Tweets and users. Possible connections are both direct and diffuse, and the hashtag and @-mention have been of special interest to linguists. According to Zappavigna (2011, 2012), relations are created when users include platform-specific tags (hashtags, @-tags) to create "searchable talk" in their Tweets. Although the platform is not unique among computer-mediated discourse for being easily searchable by keyword, the use of these tags makes searchability social in Twitter. In the case of the hashtag, users create "ambient affiliation" between disparate texts including the tag, forming a cloud of affiliation based on content rather than one-to-one, user-based connections. Page (2012) writes that hashtags project potential interaction and enable visibility through the search connections they create. Page's article also informs my view of the Tweet texts from a trending topic as a co-constructed discourse by writing that hashtags create an asymmetrical, not dialogic connection, by broadcasting talk "about" rather than creating talk "with" others. In a 2009 study of the tag, Honeycutt and Herring found coherent, dyadic conversations centered around the @-tag in about a third of its uses. Tweets appear across Twitter in the order in

which they are submitted and without regard to topic, creating a lack of turn adjacency that could make coherence difficult. However, the @-tag in Tweets often sparked coherent and collaborative threads, whose content is significantly less self-focused and more addressive than Tweets that do not include an @-tag. Sousa et al 2010 explore these ties quantitatively, tying them to social networks as they observe that users with smaller networks @-mention to create social ties with other across topics, creating a dense network, while users with larger networks show somewhat more disjointed mentioning practices, mentioning users based on topic of Tweets (though the authors maintain that these topic-motivated ties are still significantly social).

Retweeting, especially, has inspired study that ranges from qualitative inquiry into computational modeling of message dissemination or user influence. boyd, Goldan, and Lotan (2010) identify retweeting as a social and conversational practice. Though the practice of retweeting has changed significantly since the time of their writing (especially since the platform now supports a native Retweet functionality; when they wrote, users manually copied and prefaced messages), the authors show that audience design is a crucial consideration in retweeting, as users attempt to spread messages to broader audiences or specifically, to their own followers. Retweets also served to open a line of commentary or conversation on the content of a Tweet or to validate or publicly agree with another user. By examining which types of third person portrayals are retweeted, I observe the shareability of messages. However, a retweeted text suggests that it's not specific to the original author, but rather has broader applicability and resonance, suggesting less personalization. Using quantitative models, several authors have attempted to uncover why certain messages spread or to define influence within the platform. Wu et al (2011) expose a great imbalance in the platform: in asking "who Tweets

what to whom?" uncover not a comprehensive view of Tweets, but the finding that .05% of users attract 50% of attention, mostly after their Tweets have been passed through intermediary users. The authors prefer the term "information sharing network" to "social network" because of the pervasiveness non-reciprocal ties. Cha et al (2010), noting that tags, as directed links, can display multiple stances, nevertheless develop a taxonomy of tag-based influence: influence measured on frequency of Retweets shows that the content is influential, while mentions of a user betray that user's name value. The authors give agency to Twitter users, stating that influence is won only through Twitter users' personal effort and involvement, focusing on a single topic to concentrate influence. They also note that influence comes only from the susceptibility of society. Leavitt et al (2009) also distinguish between conversational (@-mention-based) and content (Retweet-based) influence, and add a conception of this movement of Tweets/user links as a social action. Suh et al (2010), looking to create a program that could engineer Retweets, look for correlation like the number of past Tweets by the author.

These authors often look at both social network context and content features of Tweets to determine retweetability. Hashtags and URLS are found by several to significantly increase a Tweet's likelihood of propagation (Suh et al 2010, Naveed et al 2011). The concept of "interestingness" is advanced by a few authors, who calculate this based on similarity of words in retweets to words in a user's original Tweets (Yang et al, 2010, showing a correlation), and a cluster of features like interpersonal and popular topic, question marks, and negativity (Naveed et al, 2011). Naveed et al are, in fact, able to model the probability of Retweeting based solely on the content of Tweets without using any social network data. They determine that "a tweet

is likely to be retweeted when it is about a general, public topic instead of a narrow, personal topic," or goes so far as to be addressive.

In this tradition, my study eventually turns to the propagation of messages within Twitter. However, this paper approaches Twitter through textual data with only limited information disclosed about social networks. That is, only connections (@-mention user-to-user connections; shared hashtags connecting texts) explicitly performed in the text are available for analysis, in comparison to the collection techniques of many other researches that crawls user networks. My collection technique, though lacking the depth of data of those studies, offers a perspective of the "surface level" connections that are easily viewable on the platform. For most users, the connections of the other, geographically disparate participants in the hashtag are unknowable and obscure, while the connections made in their texts are the primary experience of the stream. Here, I approach the question from linguistic and social premises, specifically asking what insight the propagation I find offers into linguistic form, a topic left largely unexplored in these studies.

iii. Approaches to correlating pronouns with extralinguistic features

The following study emerges from the basic premise that the patterns of alternation of linguistic items offer insight into extra-semantic information about them. In other words, it will attempt to locate a pattern around its pronoun variables in order to analyze it for contributions to the meaning of the variable. This tenet is adapted from several linguistic subfields to fit the scope and type of data analyzed here. The following review first introduces principles about the denotational value of words (including pronouns) and then moves on to two larger-scale,

quantitative approaches, one from corpus pragmatics and one from sociolinguistic variation, that guide my approach to the volume of data in my Twitter dataset.

This search for extra information in terms of reference, which fuel small-scale, qualitative studies of identity and positioning, has also been applied to larger, corpora-based studies. Brown and Gilman's seminal 1960 crosslinguistic study explored the tendencies of second-person pronouns. Based on questionnaire data, the authors examined variation in second person, "T" and "V" pronouns (so named after the French "tu" and "vous"). This correlated with "objective relationships" between speaker and hearer, and the authors found that larger patterns in the use of one pronoun or the other could be used to create speakers' "expressive styles." The authors named T and V pronouns after the relationships they projected: (respectively) the "pronouns of power and solidarity."

Later, further research emerged in the style of sociolinguistic variation that tied words' meaning to the tendencies in surrounding, extralinguistic information. Variationist study is rooted in the study of strictly semantically equivalent variants like allophones. However, its objects of study have been rethought, the principles underlying the field found to be fruitful in the study of not only semantically equivalent variants, but of differences in meaning, and of not only macro-demographic categories but also of the creation of social distinction. Since at least Lavandera's 1978 essay, the implications of expanding the scope of the sociolinguistic variable have been considered, observing that the "tendencies and frequencies" of many levels of linguistic variation may carry social meaning. Lavandera suggests a standard of "functional comparability" for variants rather than semantic equivalence. Increasingly, variationist methods have been used to analyze variants' meanings in practice rather than to correlate different

realizations of the variable with distinct demographic groups. This approach to variation in practice especially applies to the online variation where macro-demographics do not have the same salience as offline, and are certainly not always knowable. It seems especially applicable to a study of third-person singular pronouns, which formally have the same truth conditions (and different pragmatic felicity conditions), but are prima facie used differently in discourse. Several previous, variationist studies have successfully uncovered associations of their variables by studying their contexts. For example, in their 1995 study of constructed dialogue, Ferrara and Bell analyzed patterns of "BE + like" in corpora of spoken narratives as compared to other variants like "say." The differing correlations with person and advancement of the narrative show that “BE + like” is linked to internal states. In 2004, the relationships between speaker and hearer in Keisling’s study of the word “dude” in fraternity men’s speech and in a corpus collected by his students showed the word to be associated with a stance of “cool solidarity.” Finally, in a 2010 paper reviewing the study of variation in larger units of language, Pichler puts it succinctly: "in some cases, function may even exert a more important constraint on discourse variability than social factors" (597).

Indexical, social meaning is also increasingly sought in variationist studies. Third-wave variationst work now seeks to situate variants’ meaning in their social context; meaning and sociality are seen as inseparable. Arguing that meaning – including meaning of variables – is emergent in interaction, Eckert (2008) identifies the study of local social meaning and indexicalities of variants as the ultimate goal of variationist study. In a 2012 essay, Eckert further defines this ‘third wave’ of variationist research. This wave focuses on how combinations of linguistic features create meaning in interaction, “foreground(ing) the relation

between language use and the kinds of social moves that lead to the inscription of new categories and social meanings (95).” This study, once lead through a simple examination of patterns in tokens, looks to the corpus level to the inscription of meaning by the textual practices that constitute the corpus. In this way, this paper draws from the premise that corpora offer insight into of the emergence of meaning in discourse.

I also draw upon other corpus-based approaches, especially those used to examine pragmatic content. This approach to meaning in use has been studied increasingly in computational studies, which approach semantic and pragmatic questions in the tendencies of large corpora. Word meanings are taken as coming partially from the words’ collocation with other types of words and phrases, for example, as Stubbs (2001) argues, “our knowledge of a language is not only a knowledge of individual words, but of their predictable combinations, and of the cultural knowledge which these combinations often encapsulate” (3). This has, in at least one case, meant correlating extralinguistic features with target words to discern those words’ force: In a 2009 paper, Constant et al explore the use of expressives based on their use in Amazon reviews. The term expressives, describes words with, among other properties, pragmatic content about emotional states independent of truth values (cf Potts 2007), like “damn” and “bastard.” The authors of the 2009 paper correlate the frequency of several such words with the extralinguistic feature of level and valence of emotion in a corpus of Amazon.com reviews, as indicated by the number of stars in each review where an expressive is used. These tendencies allow for a more nuanced understanding of conditions of use and of the meaning expressives contribute to the texts in which they’re found. Though pronouns are not addressed in the 2009, corpus-based paper, one co-author, Potts, writes about their expressive

properties in a 2007 paper. Reminiscent of the assertions made in above, sociolinguistic citations, Potts writes that second person T and V pronouns create additional expressive meanings in addition to propositional content: “the expressive setting — the indicated relationship between speaker and addressee — is different”. This thesis, in examining the content about relationships created by the Retweet and @-tag, draws upon the work of this corpus-based approach to expressive meaning and expands it to the area of pronouns. In this study of pronouns, I attempt to apply this principle to the tendencies of THEY, with a similar attention to how features of the corpus above the level of individual texts offer insight into elements of those texts.

IV. Data

The following section describes the collection of a Twitter corpus collected based on a common, referent-introducing hashtag, which is analyzed around the third-person singular pronouns linked to that hashtag. It proceeds by detailing the collection techniques of the initial corpus of 9031 Tweets, then providing an introduction to the sort of referent constituted in the hashtag all Tweets have in common. Then, the annotation of Tweets by third-person pronoun is discussed and the methods of further, token-level annotation used in previous studies are shown to be poorly applicable to this data. Finally, I turn to a view of the texts as interrelated, clustering the corpus based on “core” texts within Tweets, thereby finding a clear distinction in the scale of dissemination of several Tweets that include THEY.

i. Collecting the #oomf-corpus

The corpus for this study was collected on February 5, 2013, between 6:54 and 7:21 PM EST, after I saw #oomf in the US Trends box while using the platform. Using ten calls to the Twitter Search API with the search term #oomf, which returns Tweets including that precise string in the Tweet text, over 10,000 public Tweets were collected. The corpus was filtered for duplicates using Tweet-specific id numbers, leaving 9031 unique Tweets to comprise the corpus, which the timestamps indicate were all tweeted within 32 minutes of each other.

The primary unit of analysis for this research is the Tweet text (that is, the brief textual strings submitted by users when they tweet) of those 9031 Tweets. Each of those tokens constitutes a unique action and a unique object within the platform and is open equally for propagation and interaction. This social uniqueness is not reflected on a textual level, however, and many of the strings making up Tweet texts are non-unique: the 9031 tokens are found to represent 6001 "types" as defined by complete string identity (that is, the Tweet texts being precisely the same combination of characters; see section *V*-iii. Sub-corpus composition for a continued discussion of how similarity between Tweet texts is analyzed in this research). The 9031 Tweets originate from 8031 unique accounts, meaning that users averaged 1.1 Tweets each in the corpus. Querying the language codes attached to Tweets indicates that 8546 of them, or nearly 95%, are in English. However, as the language code is attached to Tweets based on the users' interface language and not individual Tweet texts, my informal survey of the corpus suggests that the share of Tweets in English is actually quite close to 100% (perhaps because #oomf is an English-language abbreviation, see next section). At any rate, as this is a study of English-

language pronouns in Tweet texts, only Tweet texts determined to include English pronouns in English texts were identified for further analysis in the next phases.

This corpus offers a sample of #oomf-including, public Tweets produced during a short time period. It serves as a sample of naturally occurring data, but it is not a complete set of Tweets using the hashtag, due to several factors in the sampling. Both the timing of the sampling and Twitter's Search API itself make this an incomplete set. First, Tweets from short periods during the collection may have been missed, based on when the calls were sent to the Search API (that is, if there was a gap between the last Tweet in one batch of results and the first Tweet in the next). Second, Twitter's Search API does not use any authentication and therefore indexes only public tweets. Some users' Tweets are viewable only to followers who they have approved, but none of those are included in the present dataset. Third, the results that the Search API returns are "focused in relevance and not completeness" (Twitter n.d., Twitter 2012a). The site is vague about what parameters it uses to include (or exclude) certain results in its search, though it refers explicitly to SPAM filters. Still, there is no reason to believe that any of these limitations on the corpus affects the pattern of pronouns found in the Tweets collected, and much less that they would affect different pronouns unequally. I conceptualize the corpus as a collection of naturally occurring language in which I can explore possibilities and tendencies emerging from a comparable context.

ii. Introducing #oomf, a Twitter-specific antecedent

The corpus was collected to bring together texts about referents introduced with the hashtag #oomf. Such a referent, in this environment, is specific and is not gender-marked: #oomf stands for "one of my followers" but conventions of use restrict it to definite followers who are not

named. The hashtag has a meaning parallel to the English meaning of the tag, which develops out of users' affiliation with the trend. As explored in the literature review, hashtags are affiliative; this paper's corpus captures Tweets that overwhelmingly follow implicit conventions in meaning and tag position. This concerted use of #oomf allows me to treat the sort of referents introduced by #oomf as of a certain type unless otherwise indicated, where individual Tweets are often too short to include disambiguating context in and of themselves. By then studying pronouns linked to #oomf as an antecedent, it is possible to hold that aspect of the referents studied constant[1].

#oomf is a Twitter acronym that has been conventionalized to introduce a specific referent. It abbreviates "one of my followers," but the hashtag has a more specialized use than the English phrase. As discussed in the introduction to this thesis, Twitter users asymmetrically create connections to each other by subscribing to other users' Tweets or, in Twitter terms, "following" other users (becoming other users' "followers"). Each Tweet by a user who is being followed is included in the followers' streams, where the follower may see and interact with them. The tag #oomf is used in order to talk about one of these followers who will see the Tweet. Since followers are, by definition, Twitter users, tweeting about #oomf represents a choice to refer to the follower in a way that becomes part of a larger, shallow network of Tweets including the hashtag; unlike using an @-mention to refer to a follower, it does not create direct connections between users.

[1] In the following, I borrow Lagunoff (1997)'s practice of using the term "antecedents" to refer to both actual antecedents co-referent with and quantifiers bound to pronouns. Though this is a simplification, the processes that link these "antecedents" with the pronouns studied here are not the focus of this study.

#oomf is not completely equivalent with “one of my followers.” The English phrase "one of my followers," has non-specific uses, meaning roughly "any of my followers" as in "One of my followers has my phone, so let me know if it’s you", or “an indeterminate one of my followers” as in “One of my followers is lying to me, and I am going to be so angry when I find out who.” The Twitter hashtag, in comparison, is conventionalized for use in only the specific sense of the phrase (meaning roughly, “a specific one of my followers”, as in the Tweet, “#oomf is such a flirt, and he knows it.”). On TagDef.com, a website where users offer and vote on the accuracy of definitions for Twitter hashtags, one user-submitted entry from March 2011 defines #oomf as “talking about someone without using their real name. so you use #oomf so nobody knows the persons actual name. EX. "i really like #oomf"”. Buzzfeed, a site that aggregates "best of" lists of internet artifacts like pictures and YouTube videos, devotes a November 2012 article to #oomf after it spiked in popularity over the previous days. Tweeting with #oomf is defined as "hashtagged, institutionalized subtweeting" (that is, the practice of 'subliminal tweeting,' or Tweeting about another person without using that person’s name or username name or username) (Herman 2012). These sources also hint that #oomf is usually used within a sentence rather than outside of it to tag the Tweet. To this, I add five of my own observations based on the corpus analyzed in this study.

1. #oomf is generally used within a sentence, and may occur in any position, including to introduce a possessive (both as #oomf with no possessive marking and as #oomf’s; #oomfs is not indexed in the same search). A small percentage of Tweets in my corpus do present counterexamples to this, tagging the sentence as being about #oomf instead (e).

a) I can't believe anything #oomf says anymore.

b) I care about #Oomf to much ...

c) Got a text this morning from #oomf :)

d) RT @USER[2]: #Oomf boyfriend always tweeting cute shit to her n' shit.. must be nice.

e) #explaintomewhy he is not my boyfriend #oomf

2. #oomf is identifiable. As many #oomf Tweets play with ambiguity and obscurity, this process of identification tends to emerge as other users Retweet the original message and try to add specificity. In a, the user first begins by quoting the username and message from a previous Tweet, and the added text at the end indicates that this user believes they have been "called out", or identified in a negative context. b begins in the same way, but this time the added message urges the quoted user to @-mention #oomf (go ahead and @[-mention] 'em) - that is, explicitly identify, by linking to, the follower being discussed. In c, the user states that they know who #oomf in the text they are retweeting. In d, a user complains about the trend, saying that instead of using the hashtag, users should identify the user with a tag because it's essentially a private conversation.

a) "@USER_A: #oomf is a bitch". Way to subtly call me out!

b) "@USER_B: I wanna do sum thangs to #oomf" gon head at @ em

c) "@USER_C: Damn #oomf looks busted as hell in their avi" I was thinking the same thing cause I know who you're talking about haha

d) I hate the hashtag #oomf. if you're gonna say something only THEY know just freaking tag them. You're not being clever, you're being dumb.

[2] Though all Tweets included here are public, out of the respect for the privacy of users, I anonymize mentioned usernames here. Where multiple usernames are present in the same series of examples, I assign unique usernames letters: USER_A, etc.

3. The assertion that #oomf introduces a specific follower can be confirmed in some but not every text. Especially with the brevity of Twitter texts, many individual Tweets like a and b are ambiguous; however, the presence of Tweets like c and d urge a reading, in this context, of specific referents. A small minority of counterexamples do use #oomf non-specifically. In e, the user talks about #oomf non-specifically, but rather as a "type" of individual - the use of "you" strongly implies that the Tweet is asking if "you ever get the feeling that one of *your* followers retweets you ...".

a) #oomf looked pretty cute today

b) Well lets text #Oomf >

c) #oomf is such a flirt, and he knows it.

d) #oomf hasn't texted me all day...nice move.

e) You ever get the feeling that #oomf retweets you because he/she is trying to you to notice him/her? #TakeTheHint!

4. #oomf Tweets seem not to be restricted to specific topics in principle, but three major themes emerge: sexuality and relationships, communication with #oomf, and negativity (see Naveed et al 2011 for a discussion of the increased spread of negative Tweets on Twitter). Many Tweets, like a, are sexually explicit or discuss relationship aspirations or regrets. Many other Tweets refer to texting, talking, or Tweeting, like b, or are used to insult #oomf, like c.

a) #Oomf'sSexyAss >>>>>> lol

b) #oomf need to text me back

c) #oomf turned fake on me though lol.

5. The corpus includes many symbolic and other non-standard features. As stated in the introduction to the Twitter platform, collecting data via a hashtag results in a decidedly non random sample of users and texts that is unlikely to be representative of users in general. Rather, a few general observations to characterize the language of the corpus are offered here. First, many Tweets show AAVE and general non-standard features, for example, the progressive aspect ("steady") in a and b shows copula deletion. Second, many Tweets include visual information in their texts, too. In addition to emoticons, these include Unicode glyphs, which one-character smiley faces and hearts. These are returned from the Twitter API with their unique code. These display on some users' interfaces as smileys, pictures, etc.; for other users (my own experience on a web browser included), most of these display as boxes or empty spaces because their device or app does not support an inclusive-enough Unicode font. They are typically inserted by copying and pasting the glyph from a website offering an index of Unicode symbols. Tweet c below includes the Unicode glyph for a pensive face (represented, as collected from the API, as "\U0001f614"). And finally, Tweets often include additional internet abbreviations or modes of expression. Finally, many Tweets use internet slang or abbreviations, for example, d uses "AF" for "as fuck".

a) #oomf steady lying!

b) I wonder if #Oomf single

c) I haven't talk to #oomf all day 😔 where's my baby ?

d) #oomf is attractive AF.

The emergence of the first three commonalities over the corpus suggests a specialized use of the hashtag to introduce a specific referent. The fourth and fifth, regarding common

topics and textual varieties used in the corpus, are presented in order to characterize the corpus and put the proceeding examples in context.

V. Results

The following section describes the data and presents results in three steps. First, the corpus is annotated for target pronouns, third person pronouns co-referent with #oomf. Then, annotation of Tweets at the token level is found to be uncompelling for showing differences between the pronoun groups. Finally, Tweets are grouped into "clusters" based on textual similarities.

i. Third person, singular pronouns linked to #oomf

The Tweet texts of this corpus were annotated for presence of third person pronouns linked to #oomf. Other third person pronouns were present in the corpus, but are not factored into the analysis. Target pronouns were annotated using a Python script and my own, individualized judgments regarding Tweets. The script uses a collection of regular expressions to comb each Tweet text for potential iterations of he- she- and they-type pronouns (for example, one expression for HE searched for "he'll", ignoring character case, proceeded by and followed by non-alphanumeric characters, and with or without the apostrophe). Then, the Tweets found to contain a potential pronoun were passed to me, based on pronoun group, for input. I determined if the pronoun was valid, for example, several instances of "hell" found by the aforementioned expression were not pronouns. I also determined if the pronoun was linked to #oomf based on my reading; where the pronoun was found to have a different referent, it was separately annotated and not factored into the analysis here. This method of annotation does

limit the pronouns found to HE, SHE and THEY-type pronouns; if other third-person singular pronouns, such as hir or ze are present, they are not considered here. Table 1: Strings found as target pronouns in corpus, by pronoun groupbelow presents the strings determined in at least one Tweet in this corpus to be pronouns linked to #oomf; note that this was shaped heavily by the actuality of the corpus and they are neither all standard nor an exhaustive set of all pronouns (for example, no Tweet was found to include "she'd").

group	# of 1139	found in corpus as
THEY	379	them, there, their, themself, they, themselves, them self
SHE	407	she'll, her, she, hers, she's, herself, shes
HE	364	his, hes, he's, he'll, he'd, himself, him, he, hisself

Table 1: Strings found as target pronouns in corpus, by pronoun group

These target pronouns are found in 1139 Tweets in the corpus, or about 12.6% of the 9031 Tweets, in an approximately equal distribution between SHE, HE and THEY.

Figure 2, below, illustrates the quantity of pronouns in each group. Of Tweets containing a target pronoun, approximately 36% contain SHE, 31% HE and 32% THEY[3]. This balance is, indeed, unique among the studies I've read: most sources report that male referents are vastly overrepresented in discourse (for example, Balhorn 2009 reports a much higher incidence of HE than SHE or THEY in his newspaper corpus; Twenge et al 2012 survey books ranging from 1990 to 2008 and find masculine referents twice as prevalent as feminine). Elicitation- and corpus-based of generics also seldom find data balanced between SHE, HE and THEY, almost always showing few generic uses of SHE (see, for example, Myers 1990). In this corpus, however,

[3] Eleven of these Tweets use two pronoun groups, a few as a he/she type indefinite, a few to talk about the relationship between a SHE-#oomf and a HE-#oomf, and one to refer to the same #oomf as both SHE and THEY.

portrayals of the target referent with each of the three pronoun groups occur at approximately the same rate. This has the immediate consequence of simplifying the current study by allowing the subcorpora to be compared to one another more directly. Additional considerations of the balance of gendered (and non-gendered) pronouns in current corpora or in corpora with less restricted participation (as compared to Balhorn's newspaper prose, Twenge's published books) could serve as a direction of future research.

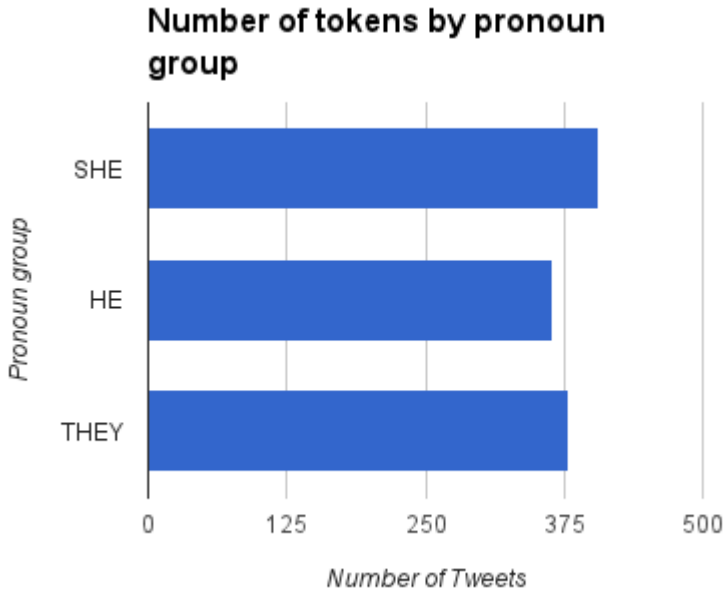

Figure 2: Number of Tweets included in each pronoun group

ii. Tweet-level variation

The first level of inquiry into the corpus, at the Tweet level, poorly accounts for the differences between pronouns data studied in this level, suggesting instead similarity in referent or inadequacy of methods. Variation at the level of tokens is examined in referent features (based on variables in previous corpus-based THEY studies) and in the presence of "minimal pair"-like Tweets that show the target pronouns being used in nearly identical Tweets. These tests seem to indicate similarity in the tokens, rather than patterning with variation. Then, the strategy of

measuring co-occurrence of pronouns with additional qualitative features of Tweets (ex. Retweet tags) is rejected, as it is found to poorly represent this data.

a. Lack of variation in referent

Previous authors focused on how features of the referent varied between THEY, SHE and HE; Tweets were exploratively but ultimately, uncompellingly tested for several of these features when separable from antecedent. Several of the previous studies reviewed above deal narrowly with the combination of pronouns and antecedents, distinguishing between, for example, quantifiers (gender-marked or not), generic definites, existential indefinites, and (unattested with THEY) names and deixis in Lagunoff (1997). Though each author offers a slightly different taxonomy of possible antecedents, it is clear that none will account for the tokens included in this analysis, which, definitionally, are all linked to the same antecedent (#oomf). Furthermore, as established above, in the vast majority of Tweets, #oomf is a specific noun-phrase antecedent - what Lagunoff (1997) and Newman (1992) each claim to be the least likely of the possible antecedents to be linked to THEY. Still, this assertion of unlikelihood poorly accounts for this naturally occurring corpus, in which THEY occurs at nearly the same number of times and SHE or HE.

Several features proposed independently of antecedents were tested on a small, random sample of Tweet texts to confirm that, in practice, these features did not compellingly correlate with pronoun variation within this corpus. I examine gender marking in the Tweets alongside several referent features analyzed in previous studies: notional number, referential solidity, and individuation. Notional number (Newman 1992, Baranowski 2002) allows for the possibility of multiple people as the referent (for example, ‘Every #oomf tweeted about their

day,' if it were in the corpus, could be judged to be notionally plural). Referential solidity (Newman 1992) tries to capture the idea that a referent may or may not actually exist, such as in the fictional Tweet 'I wish I had #oomf that would Tweet about his day'. Individuation, which also plays an important role in the observations of introspective studies like McConnell-Ginet 2011a, 2011b and Mackay 1980, is the concept that some people enter discourse as prototypes and not as specific characters; Newman (1998) quantifies it on a scale from 1 (least individuated) to 5 (most individuated), but I find this degree of distinction untenable for such short texts and reduce it to 1-3. The concept of gender applied here is my own: attempting to recognize that looking for gender only in the pronoun would beg the question, I look at all other aspects of the sentence for gendering of the referent. I thereby recognize that gender is *not* a salient feature outside of antecedents and pronouns in most sentences and finding THEY in sentences without additional gender information does not prove a unique point about THEY. In practice, I find gendering in my samples with words like "bro" and "bitch" though I looked in principle for any overtly gendered NPs, modifiers, or predicates associated with #oomf. A table defining these measures and tables showing them applied to SHE, HE, and THEY are included in the four tables of the appendix.

This small, random sample serves to illustrate that the features associated with THEY in other studies neither apply well to THEY in these cases, nor do they vary between the pronoun groups in this corpus. Each pronoun group had other gender information in only one of ten Tweets. The #oomf in every token was plausibly singular and solid, referring often to past and ongoing interactions with specific people or to physical appearance. Because of this level of specificity, all samples except one THEY Tweet are judged to be fully individuated. It seems that

all these traits approached the idea of 'specificity' of the referent but that in this corpus, which, with this antecedent, shows little difference across pronoun groups. As previously examined aspects of the referent fail to motivate distinctions between THEY, SHE and HE, this small sample provides strong evidence that the salient and compelling distinctions between the target pronouns in this corpus are to be found elsewhere.

b. Cross-pronoun Tweets

Indeed, the similarities between pronoun-including Tweets extend beyond conceptualizations of the referent, but are written also into the entire Tweet texts. A close examination of the pronoun subcorpora reveals that several Tweets cross between corpora. While some of these Tweets seem to be derivations of each other, others seem to be rather mundane messages that were simply expressed by several users, linking different pronouns to #oomf. Inconsistencies among similar messages' typography (ex spelling, punctuation) and sentence structure are strong evidence for or against shared origins (or at least a desire to display shared origins) in this context, where messages are so easily and so often copied and pasted. In the following, several sets of Tweets illustrate that THEY can alternate with SHE or HE both incidentally and, apparently, through intentional replacement.

These Tweets are something of minimal pairs between the pronouns, supporting the hypothesis that it is the pronouns themselves that contrast with each other in the sentence and suggesting that attention be paid to the context rather than the sentence. A sentence with SHE is not distinguished from a sentence with HE by the propositional content of the sentence, but rather by the appropriateness of the reference in the context (cf Heim and Kratzer 1998). For example, between in the following two pairs of Tweets, the contrast would be described in

terms of felicity conditions: the sentence is meaningful if and only if the features of the pronoun, to begin with, that of gender, match those of the referent. Linguists, when describing these, would not look to the minor differences between them (like the replacement of “does” and “if”) to explain this difference.

- I Wonder Does #Oomf Ever Think About Me Like I Think About Him..?
- I Wonder If #Oomf Thinks About Me, Like I Think About Her?

Similar pairs are found between THEY and HE, once again with no propositional content of the sentence offering clues about what distinguishes the two pronouns from each other. These especially question how the features of the referent that were examined above in random tokens could be applied on a token level: if THEY does imply a less individuated or notionally plural referent, that would not be discernable from these sentences without begging the question.

- hoping #oomf sees all these tweets about them....
- I hope #oomf sees my last tweet and knows its about him.

And:

- #oomf going through. Poor baby. He needs a hug..
- #oomf going through hope they gone be alright

[Note: ‘going through’ means roughly ‘going through a difficult time’]

One particular predicate is used with #oomf in Tweets about SHE, HE, and THEY. These Tweets assert that #oomf “could get it,” meaning “could have sex with somebody” (usually meaning the speaker, where not further specified). This is a common comment about #oomf in the corpus (including in several Tweets that did not include target pronouns). Apparently, a

pronoun must not encode information about the referent's gender to be used with this sexual predicate. In the following, "can get it" is used with all three pronoun groups, in a and b with THEY, in c with HE, and in d-f with SHE.

a) Yeeeeessss!! RT "@USER_A: #oomf just don't know how bad they could get it"

b) #oomf just don't know how bad they could get it

c) RT @USER_B: If #oomf hadn't fkd my friend he could get it 😒□□✋

[Note: fkd abbreviates "fucked"]

d) Omg yoo #oomf could get it lol she hella right too

e) #oomf know she can get it ;)

f) #oomf can get it and she knows it

Finally, one minimal-pair-like set of Tweets seems to have emerged with intentional replacement of THEY with SHE. In this example, extrapolating from the Tweets' timestamps, unlikely message, and shared typographical features of the capitalization of Avi (for 'avatar,' not usually capitalized) and ampersand, a THEY Tweet is copied and changed in several places (pronoun, expressive, addition of a Unicode character). Each of these iterations is then Retweeted by several additional users (seven times for a, twelve times for b).

a) #OOMF black as HELL! I clicked on their Avi & thought my phone died

b) #Oomf black AS FUCK ! I clicked her Avi & thought my phone died. 😍

These similar Tweets emerging from the corpus suggest that sentences in which THEY may be appropriate overlap with sentences in which SHE or HE are appropriate. The usefulness of looking to individual Tweets for features that condition a specific pronoun is limited by the consideration that a given context may condition multiple pronouns. Instead, they suggest

something in context, rather than within sentences, conditions the appropriateness of the pronoun used.

c. Token-level co-occurrence

Based on the methodology of the Twitter literature cited in the above sections, an attempt was made to annotate and cross-tabulate additional features of Tweets, such as the co-occurrence of the target pronouns, Twitter tags (@-mentions, Retweets, additional hashtags) and additional pronouns. This was done using a similar Python script to the one that found third-person pronouns, although these features were generally more straightforward and did not require human input. Ultimately, these results are not included in detail here because these co-occurrences were greatly skewed by a certain tendencies of the data at the token level, and the subcorpora were inconsistently reduced at the type level. For instance, the most compelling among mostly-insignificant measures was the finding that fewer of the Tweets including SHE or HE were marked as Retweets (about 32% and 25% of each subcorpus, respectively) than the corpus in general (about 54%), while a larger share of THEY Tweets (64%) were Retweets. These correlations could be read similarly to the results that motivated, among others, Naveed et al's (2010), Suh et al's (2010) and Yang et al's (2010) accounts of what makes a Tweet more retweetable. However, to me, it seems too strong a claim to make that a pronoun sparks retweeting behaviors or even to suggest more moderately statements about the sharability of the Tweets' content. Rather, based on closer examination of the subcorpora, these likelihoods of retweeting were determined by a few aberrant cases, while most THEY Tweets were spread quite similarly to the Tweets in the other subcorpora. That is, upon closer examination of these subcorpora, it did not seem to be the case that each Tweet enjoyed equal

chances at interaction, by virtue of containing a certain pronoun. Instead, the pattern emerged on a corpus level once connections *between* heavily retweeted texts were established.

iii. Sub-corpus composition

The following section introduces another method of analysis of this corpus: clustering Tweets based on shared texts at their cores. In what follows, I ground my method of clustering Tweets around what I call "cores" (recognizably shared texts) and then describe the results of this process, which reveal a very different composition of the THEY subcorpus than of the other subcorpora.

a. "Core texts" and clustering

Pronouns are features of individual texts, and Tweets the basic unit of interaction in Twitter; however, textual and social connections are performed within those texts and nuanced view of the linguistics of Twitter demands also an ethnographic eye to those connections and affiliations. Though the subcorpus of target-pronoun-including Tweets studied here contains 1139 unique Tweets and unique points of social action, those 1139 do not represent as unconnected, independent data points. In the following sections, I reject a view of this corpus as a two-dimensional stream of texts and instead associate the Tweets based on textual similarities.

The most basic and pervasive connection in this hashtag corpus is the act of reproducing another user's text. Here, I look for more than paraphrase or shared sentiment; instead, I focus on clearly related Tweets that share a core text: what was apparently originally a bare Tweet and is now central to one or more Tweets, perhaps surrounded by Retweet markings and additional commentary. This goes further than reducing the corpus from tokens to "types"

based on string identity of the Tweet texts: such a narrow definition of sameness captures some connections but misses some of the most basic, finding, for example, the two iterations by separate users, with unique timestamps, of "'RT @USER: I Love #oomf And I Told Her" but not the original version that is also present in the corpus, "I Love #oomf and I Told Her". Instead, by clustering texts that apparently arise as the same "core" text is modified and appended, connected texts are brought together. Where certain clusters did not have a discoverable, "core" Tweet without any Retweet markings and the earliest timestamp, one could be inferred.

This method exposes the intertextual and interpersonal connections written into the corpus, but is closely tied to the time of sampling and is certainly a situated, partial account of connections made with the texts studied. For example, several Tweets that did not seem related to any other Tweets in the corpus were, nevertheless, marked as Retweets. This may be due to the limited timeframe of my sampling of the API or may reflect incomplete datasets. It is impossible to know, also, which Tweets will later be retweeted or otherwise shared, which were not broadly shared at the time of collection. This points, more than anything, to the incomplete nature of my study and reflects a general principle of social science: that any study is necessarily a partial view of any entity from the perspective of a unique situation in time as space (cf Bucholtz and Hall 2005's "partialness principle"). Still, the emergence of interaction and scales of interaction in such a limited dataset does seem meaningful.

My decision to study Tweets as "clusters"[4] is grounded in the nature of the data as texts. Other authors, borrowing from computer science literature, refer to a "cascade" model of

[4] Cheong and Lee 2010 use the term "cluster" distinctly, as they group Tweets based on patterns of Tweet attributes and user attributes together, thereby organizing the diversity of users (ex. nationality) and types of messages (ex: advertising) into feature profiles.

Tweets, using the term to describe how Tweets spread from user to user, being further retweeted from each node (referred to in, for example, Cha et al's description of influential users setting off 'cascades', 2010 p 2). This does more accurately capture the networked behavior behind Tweeting dissemination. In this paper, however, I avoid the term because my data is not nuanced enough to reliably reveal cascades. Twitter texts do not encode connections between users or texts, except those made explicit in texts through @-tagging or Retweet tags. In this way, much of the networked behavior between Tweets is neither visible to the researcher, nor to users who did not immediately take part in it. Where native Retweets are retweeted from an intermediary, rather than the original author, they are still only attributed to the original author; user-typed Retweet practices, in turn, are inconsistent in this regard. Furthermore, some clusters emerge without full or accurate attribution to an "original" source, but, nevertheless, consist of non-trivially similar texts; these seem to participate in the spread of the text while perhaps remaining ambivalent about user-to-user connections. Therefore, I ground this analysis in the data itself, which offers rich intertextual, if not interpersonal, information. By this I mean, patterns and trends are observable between texts, beyond those present in individual texts, and I now turn my analysis to those.

The corpus was clustered semi-automatically with a Python script and my input. A Python script was passed through each pronoun group subcorpus, and the combination of all pronoun-including Tweets to find close textual matches based on substring insertion and deletion. This definition of closeness offers the advantage of reflecting how Tweets move between users in Twitter: by attaching textual connections to other users (e.g. Retweets), using copy-paste, and insertion and deletion (which leaves intact anomalies like idiosyncratic

capitalization), rather than making character-level changes. The clusters were then reviewed to ensure that they seem to constitute a meaningful connection between texts; this is illustrated in the following two examples, one of texts that are clearly derived from a common source, and another that shows incidental similarities. Clusters that did not seem to reflect purposeful reproduction of other messages were split.

The following illustrates a set of Tweets clustered based on a common core Tweet (a) where the Tweets show diverging and layered Retweet practices (see differing Retweet techniques of b, which marks the Retweet with RT, and c, which puts the attribution and text in quotation marks; d is another layer of Retweet over c, where a string insertion at the end of c that is reproduced in d ("RTF" for "re-fucking-tweet"). *(Note: "hitting up" a person means "initiating communication with.")*

a) I'll stop hitting #oomf up now. For good. I won't put the effort in if they won't.
b) RT @USER_A: I'll stop hitting #oomf up now. For good. I won't put the effort in if they won't.
c) "@USER_A: I'll stop hitting #oomf up now. For good. I won't put the effort in if they won't." RFT
d) RT @USER_B: "@USER_A: I'll stop hitting #oomf up now. For good. I won't put the effort in if they won't." RFT

Clusters were manually split where they did not seem to capture common origins of the Tweets. The following Tweets, for example, seem to be clustered based on common form of, "I wanna ____ #oomf but ____ don[']t wanna ____ them … ". However, this message does not seem to produce particularly unlikely sentences or bring together unlikely typography, other RT

connections were apparently made between smaller subgroups, and the meaning of the Tweets in relation to each other does not suggest that the users were trying to connect them. Therefore, I determined these Tweets to actually represent two sets of clustered Tweets (a-c, d and e) and one independent Tweet (f). The hyphens and underscores at the end of 6 constitute a bored or neutral faced emoticon.

a) I wanna save #oomf but they dont wanna save themself <<<
b) "@USER_A: I wanna save #oomf but they dont wanna save themself <<<"
c) RT @USER_A: I wanna save #oomf but they dont wanna save themself <<<
d) I wanna text #oomf but I don't wanna text them first. I would feel awkward.
e) RT @USER_D: I wanna text #oomf but I don't wanna text them first. I would feel awkward.
f) I wanna talk to #Oomf but I don't wanna bother them -___-

This method of analysis does not suggest an increased tendency to retweet any given Tweet based on features; rather, it exposes the emergence of several large-scale THEY Tweets while all other Tweets bring together the interaction of only a few or one users. This perspective on the corpus uncovers a clear difference in the composition of the THEY subcorpus.

b. Cluster patterns

While each pronoun subcorpus showed many Tweets on a small scale, passing mostly through only one or two accounts, the THEY corpus included fewer cores in total, and it was nearly half comprised by only two core texts. These were Tweeted 102 and 76 times.

The breakdown of the number of unique cores is compared to the number of tokens, or unique Tweets, in Figure 3 below. This graph illustrates that, although each pronoun group

occurs in approximately the same number of Tweets in the corpus, there are many fewer messages that make up the THEY Tweets (149 cores) than SHE and HE. SHE (310 cores) and HE (290 cores) remain somewhat comparable, however, with roughly twice as many cores.

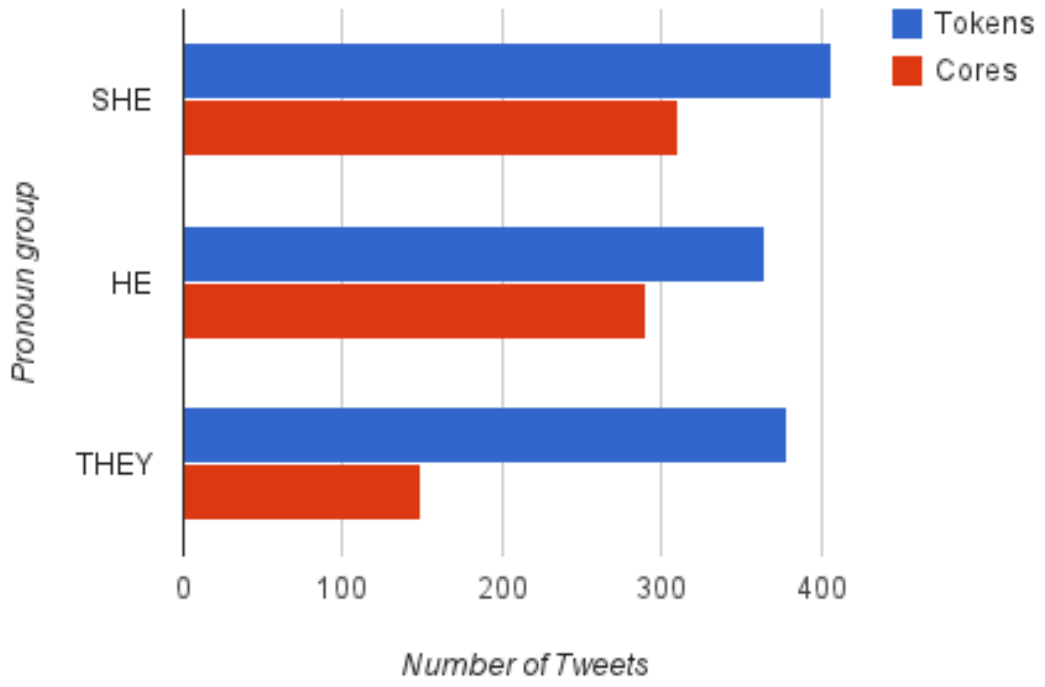

Figure 3: Number of Tokens (unique Tweets) and number of cores (unique core texts) by pronoun group.

However, it is the scale of several of the clusters that most compellingly distinguishes the THEY subcorpus from the others. All clusters of five or more Tweets are shown in Figure 4 below; this is five clusters each for SHE and HE and six for THEY. But while the largest SHE and HE clusters represent maximally 8 and 14 Tweets, respectively, two THEY clusters reach a different scale of popularity, with 102 and 76 unique Tweets constituting each cluster by virtue of being built around the same core.

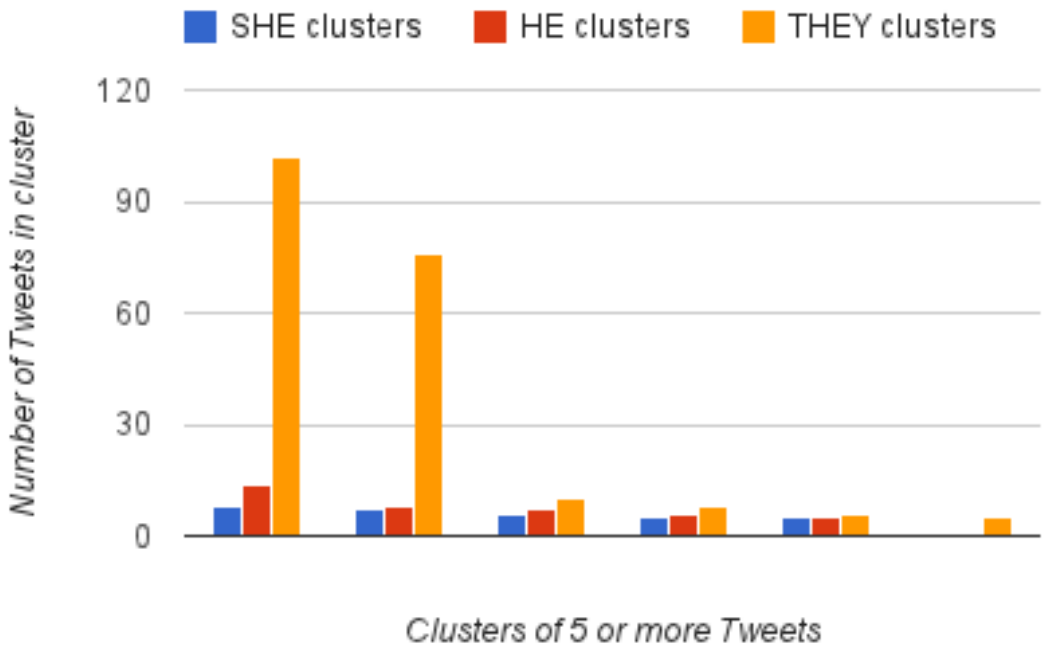

Figure 4: All clusters of 5 or more Tweets are shown, compared in size by pronoun group (THEY has one more such cluster than SHE or HE).

The vastly different scale of THEY clusters means that the subcorpora have strikingly different compositions, as illustrated in Figure 5 (SHE), Figure 6 (HE) and

Figure 7 (THEY) below. The figures show how each pronoun is tweeted with different levels of uniqueness. Clusters of four or more Tweets are represented individually, while smaller cluster are aggregated into one slice (for example, the "triples" slice in Figure 6, showing clusters in the HE subcorpus, represents five clusters of three Tweets each, and is weighted to take the space of fifteen Tweets). While the THEY subcorpus is dominated by the two large clusters, single- and two-Tweet clusters dominate the SHE and HE subcorpus. Still in each, only about ten to twelve percent of clusters reach a small-intermediate size: even THEY has a strong showing of cores that do not move between users.

Figures 5 and 6, illustrating the makeup and SHE and HE respectively, show very similar make-ups. For SHE, 63.4% of the subcorpus is made up of "singles", or clusters made up of a

core that is tweeted only once. An additional 13.8% is contained in clusters of two Tweets, and 10.3% in clusters of three Tweets. The remaining 12.5% are in 10 clusters of four to eight Tweets each. Tweets in the HE corpus were disseminated on a similar scale. 69.5% of that subcorpus is made up of "singles." 14.3% of Tweets are part of two-Tweet clusters, and 4.1% part of three-Tweet clusters. The remaining 12.1% of the corpus is included in six clusters of four to fourteen Tweets each.

For THEY, however, smaller-scale participation exists alongside massively public Tweets. 31.7% of the subcorpus is made up of cores that occur only once, 10.0% of two-Tweet clusters, and 1.6% of three-Tweet clusters. Six small clusters of between 4 and 10 Tweets up 9.7% of Tweets in the subcorpus. Finally, only two cores constitute the remaining 47% of Tweets including THEY, as they are passed from user to user: A 76-Tweet cluster represents 20.1% of the corpus and a 102-Tweet cluster, an additional 26.9%.

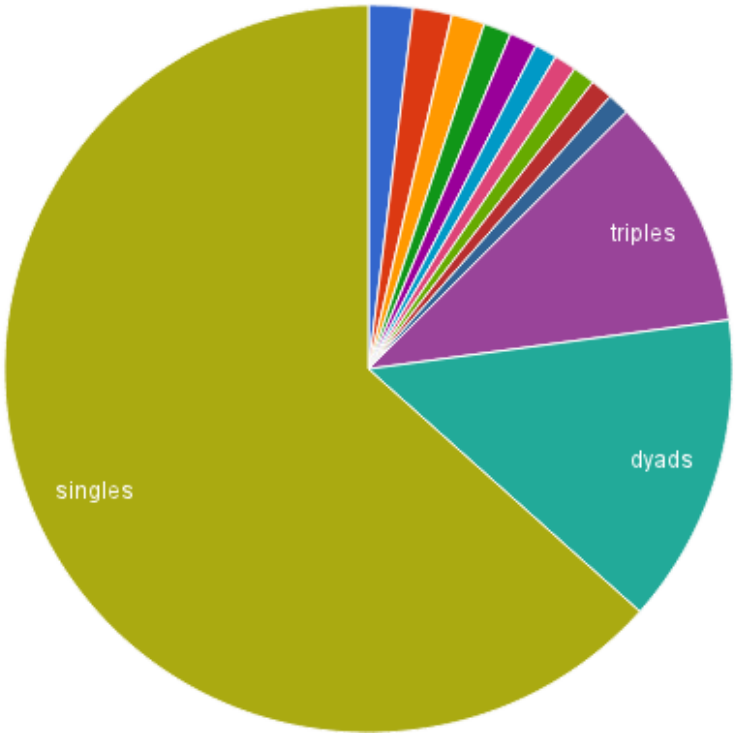

Figure 5: Composition of the 407-Tweet SHE subcorpus, by cluster size.

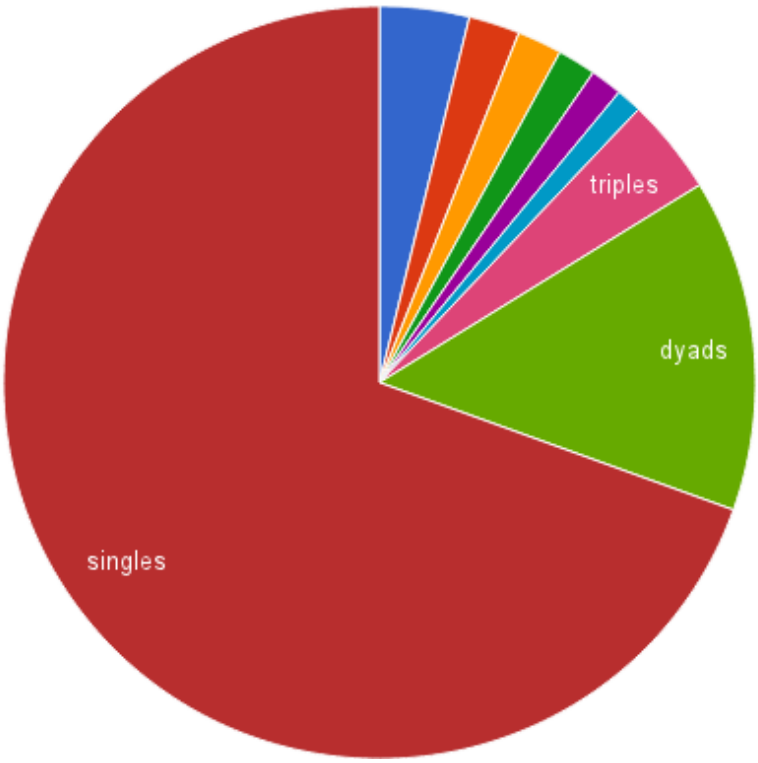

Figure 6: Composition of the 364-Tweet HE subcorpus , by cluster size.

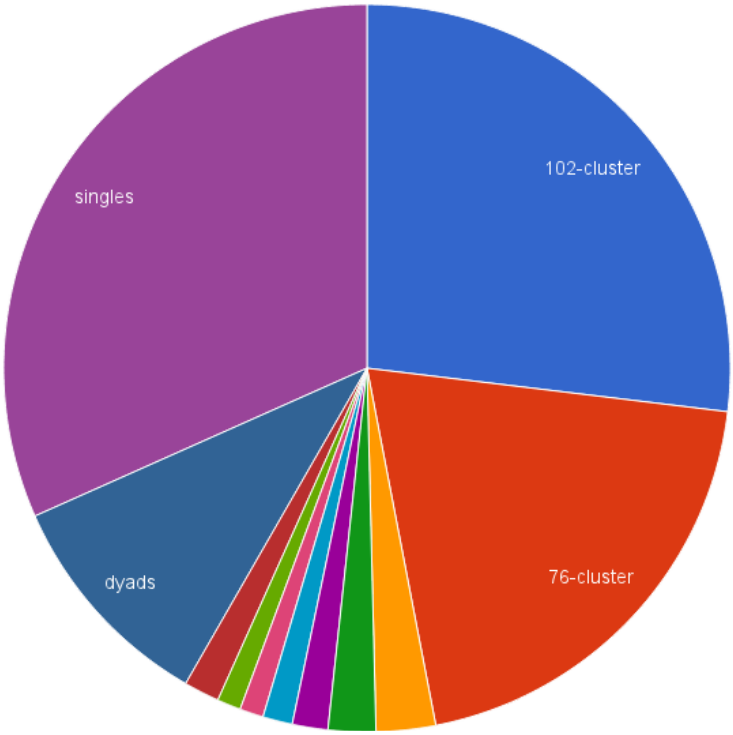

Figure 7: Composition of the 379-Tweet SHE subcorpus, by cluster size.

This means that THEY Tweets, rather than generally motivating more sharing, in two cases produced a completely different scale of text while most often being Tweeted at small scales in the same way that SHE and HE Tweets are. The two THEY cores that reached such massive popularity are reproduced below. Both apparently refer to romantic relationships between the user and #oomf.

THEY cluster 1: 102 Tweets with 9 Tweet texts. 27% of the THEY subcorpus and over 1% of the entire corpus collected. The base Tweet is as follows, including Unicode characters that illustrate each NP. The core tweet is as follows:

- #Oomf used to be my smile 😁, my world □, my heart ❤, my bestfriend □, my everything □. But now.. They ain't sh*t □.

THEY cluster 2: 76 tokens made up of 14 different types. This represents 20% of the THEY subcorpus, the basic iteration is as follows:

- If you flirt with #oomf knowing that i like them, i will personally escort you to the gates of hell.

These cores reach massive popularity, but not without additional input. Users, as they spread the core messages, append them with further specification of #oomf, further defining #oomf. The first of the following was added to the beginning of one Retweet of core 2, while b-e were added to the end of additional, apparently-retweeted versions:

a) Girl who's name starts with a N! …
b) … @USERX
c) … #girlfriendprobs
d) … A few would pass.
e) … Oh yes I would.

These engagements with the text narrow it. Especially, 1, 2, and 3 further specify the gender of the parties involved. If 1 and 2 can be interpreted as specifiers of #oomf; they would seem to indicate a lack of concern with maintaining the gender ambivalence of the core text.

These Tweets, viewed in light of the corpus as a whole, also remove any doubt that this is simply a study of linguistic form; rather it becomes clear that the spreading of texts on Twitter is social action taken at the level of Tweets, situated within online identities and structures of interaction. Upon searching the corpus for other Retweets attributed to this author, the user (an account dedicated specifically to #oomf Tweets) was found to be cited in 1476 Tweets in the corpus, or 16% of all Tweets. The mostly widely spread of these (for the

core tweet, “I would sexually destroy #oomf...”) is tweeted over 500 times with attribution to this user.

This results section has moved from a focus on Tweet features at the token level to attention to the social and intertextual connections between the texts. This level of analysis finds that the most compelling pattern in the corpus does not occur based on previous parameters of study at a token level, but rather emerges from the practice of many users sharing two THEY texts while most others are published by only one or two users.

VI. Discussion

This discussion proceeds in two veins: First, establishing the need to look beyond the token-level in this corpus, and second, proposing that the social behavior spreading two THEY messages makes it *more* public, thereby locating the features previously associated with THEY in intertextual behavior rather than individual texts.

i. Token-level similarities

Previous studies examined THEY in the context of individual utterances, looking for differences in referent type or exploring the sentence for indications of sex or sexism. Such inquiry here, holding the referent constant and examining the pronoun in alternation with, rather than opposition to, HE and SHE has shown not that THEY behaves differently from the traditional, standard pronouns, but that it is able to behave surprisingly like them. It is seen, in several Tweets in the corpus, to fill almost precisely the place in a text, being tweeted in parallel both through copy-paste and through apparently spontaneous expression of the same idea. On a sentential level, restrictions on THEY usage have not been found.

Furthermore, this has been shown for specific THEY, a type of THEY that appeared in previous studies only as an aside or an exception. Previous studies noted extremely few examples in their corpora or according to intuition (cf Balhorn 2004, 2009; Lagunoff 1997, McConnell-Ginet 2011a, 2011b; Newman 1992, 1998). However, this corpus, with 179 unique THEY sentences in 379 tokens, puts any doubt to rest that THEY is in use for known persons. This study moves away from focus of many studies on generics by showing that the use of THEY, at least in a syntactic sense, is not limited to genericness. Beyond that, it breaks from studies or popular discourse that would analyze THEY to be a pronoun developed to fill the gap of talking about people whose gender is unknown in the singular: THEY is chosen by fully one third of the Twitter users who use pronouns in this corpus, and certainly in many cases where these users have gender information available about the specific #oomf.

Furthermore, contexts with THEY overlap contexts with the standard pronouns: in antecedents, in additional features of the referent, and even at the level of entire Tweets. That the pronouns occur at roughly equal levels in this naturally occurring data with a constant antecedent motivates a look beyond antecedent, surely the most common variable studied with THEY. Previous studies either elicited pronouns with generic or non-specific referents (for example, Myers 1990, Matossian 1997) or studied variation of antecedents across naturally-occurring corpora (often focusing on generics). The approximately equal level of pronouns in this corpus was not found in any other study and is by no means the normal behavior of a corpus. While previous studies would seem to suggest #oomf is a highly unlikely type of antecedent for THEY because it is specific and definite, the conventions of use of this antecedent may make it more likely with THEY. #oomf is almost definitionally indirect, since

using a hashtag creates affiliation with other texts across the entire Twitter platform, while the available strategy of @-mentioning the follower would create a direct link between the Tweet and the mentioned user. This indirectness may be related to certain aspects of THEY written about in previous research. For instance, could indirectness be accomplished with the "low individuation" that McConnell-Ginet associates with THEY (2011b)? Though she suggests the creation of prototypes (where many THEY Tweets do describe very specific scenarios, only without naming names), perhaps simply offering fewer individuating features, like gender, for a referent can be reconciled with her account. Or, as Balhorn (2004) and Lagunoff (1997) propose in slightly different iterations, a pronoun that encodes gender may represent 'too much' information in some contexts. Though those authors wrote specifically about unknown people and generics, respectively, this concept could perhaps be extended to #oomf, who is known and specific, if it is posited that speaking about a person using gender is less direct than doing so. Still even if these considerations reconcile the relative prevalence of THEY with an apparently-unlikely antecedent type, they do little to explain why THEY occurs at precisely the level it does and why it is used in any individual text rather than SHE or HE, since presumably the motivation to be indirect applies to all Tweets equally.

To that end, measures used by previous studies to distinguish THEY-referents from those of other pronouns have been found to apply very poorly to this corpus. Based on the sampling of ten random Tweet texts for each pronoun group, this line of inquiry was not pursued. Indeed, it seemed that these features of the referent were also mostly held constant by the particular antecedent and trend studied. Importantly, gender was not an interesting variable in this brief examination: only one out of ten samples from each group contained

information about #oomf's gender within the Tweet (excepting pronouns, of course). Since #oomf is not gender-marked, and there is a general lack of further, co-occurring specifications of #oomf, many texts simply do not include gendering information about #oomf. The fact that many #oomf Tweets (though definitely not all) do not gender their referent by means outside of the pronoun is manifestly uninteresting.

The discovery of minimal-pair like Tweets and clusters also motivates a search beyond the token level for distinctions between pronouns. It establishes that there is at least an overlap, if not complete identity, of possible environments for SHE, HE and THEY. In these and, presumably, other sentences, it is not the sentence or the token that demands a pronoun with gender marking or, if it is a feature of they, a certain degree of individuation. Instead, in those sentences, it must be a pragmatic feature of the pronoun that determines appropriateness, determined in the context rather than encoded in the proposition of the sentence. Gender is not apparently not sufficient, as THEY is sometimes contrasted with the gendered pronouns, and sometimes co-occurs with gendering information about the referent. Indeed, the transformation of THEY ("OOMF black as HELL! I clicked on their Avi & thought my phone died.") into SHE ("#Oomf black AS FUCK ! I clicked her Avi & thought my phone died. 😍") especially suggests that THEY contrasts with the other pronouns: Although THEY is often analyzed as 'gender neutral,' 'epicene' or 'gender-inclusive,' its transformation into SHE shows that at least one user did not find THEY appropriate for a feminine referent. This recalls previous studies asking participants to envision a sentence with a singular they showed an extreme male bias in the descriptions and pictures produced (see Khosrashahi 1989), as well as prescriptive complaints that the pronoun is distractingly vague or impersonal (see Mackay

1980) – further cognitive studies could perhaps approach the associations unknowable from bare texts. In another case, core texts including THEY were sometimes appended with further information about #oomf that indicated gender (for instance when “Girl who's name starts with a N!” is put on the front of a Retweet of “If you flirt with #oomf knowing that i like them, i will personally escort you to the gates of hell.”) In this case, gender information about #oomf (that #oomf is a “girl”) may be made apparent while still using THEY. It seems the level of information shared about the referent’s gender, nor the level of specificity motivates the use of one pronoun over the others here.

As to the possibility of quantitatively correlating Tweet tags or other linguistic forms with tokens in the corpus individually, this seemed to poorly reflect the actual complexity of the corpus and was also not pursued; even where the raw percentage of THEY Tweets that are Retweets initially suggests that THEY is inherently more shareable, the naivety of this view is uncovered when the interconnection of tokens in the corpora are more closely examined (see next section).

Indeed, in this corpus it seems unproductive to approach variation at a sentence level between THEY, SHE and HE as it would be to ask which sentences are likely to occasion SHE over HE. The variation between SHE and HE is widely understood to be based on appropriateness of the reference taken from context; for many THEY texts that overlap the with the contents of SHE and HE Tweets, the motivation to use that pronoun must also be motivated outside the sentence. Therefore, this analysis moves to the distinctions made between pronouns on a corpus level.

ii. Corpus-level distinctions

While the results on the sentential level were characterized by similarity between the three pronouns, at a discourse level, the subcorpus of THEY Tweets showed a distinctly different composition from the other two pronoun subcorpora. Each pronoun group was almost equally represented in number of Tweets, and most of the texts in each subcorpus were shared between only a few users. Yet, while messages containing SHE and HE were reproduced in remarkably similar, invariably small-scale fashions and characterized by multiplicity of core tweets, nearly half of all THEY tweets derived from just two original texts. The token level analysis finds THEY in similar and even identical texts to SHE or HE; this corpus-level trend demands attention be the social performance of those texts. Ultimately, I argue that the features of "low individuation" and "prototypicality" previously associated with THEY emerge, in this corpus, from social behavior.

Though the THEY subcorpus is ultimately comprised of fewer core Tweets, it represents roughly the same number of unique contributions. This lack of novel linguistic information, however, does not imply an impoverished dataset: Though the core Tweet may remain untouched, by virtue of introduction into the new environment of the retweeting user, the text itself is redefined and made unique. As Becker (1984) reports of translation between Old Javanese and English, no text may ever be reproduced without producing "exuberances" and "deficiencies" of meaning. Moving words into a new environment, here into the new context of a new, unique user's account, is recognized as an action that creates additional meaning. In this case, the text appears with another users' name and/or avatar (depending on retweeting method and app used), and in that users' contexts, such as to that users' followers. It is taken

up by that user and so represented throughout the corpus. This is similar to what Bakhtin (1981) would call double-voicing (or, in light of the plurality of users in the clusters studied here, may be called polyvocality): the trace of one utterance taken up in another voice. Computers scientists write of influence by studying interpersonal connections on Twitter; Bakhtin writes of influence beyond forms of transmitting another's discourse," focusing instead on the connections in the text: "When such influences are laid bare, the half-concealed life lived by another's discourse is revealed within the new context of the given author (1981, p 347)."

These meanings created between users, rather than the smaller inventory of apparently original texts, are exuberant in the publicness they constitute for THEY. Tokens in and of themselves fail to offer enough pragmatic context to contrast the pronouns; however, their combination and textual connections in the corpus suggest more about how THEY is used. Java et al (2009) write that Twitter is a scale-free environment: Tweets and users all begin, apparently, on the same level and may be repeated and connected through the platform at any scale. The pronoun-including texts in this corpus seem to remain on one scale, except for two that include THEY, which are shared between many users and introduced into many contexts. As Baym and boyd (2012) identify layers and levels of publicness in social media, these texts seem to become *more* public. Though every Tweet collected was public and open to the same possibilities of interaction and repetition as these, most stayed between only a few users and only a few contexts. THEY, however, was brought to a different layer of publicness through the actions of many users. This publicness is defined between the many texts in the *actuality* of repetition, rather than the possibility. This publicness far exceeds the possibility that THEY is simply shared both by people tweeting about male and about female #oomfs. It far exceeds any

sort of “double sharability” (assuming most #oomfs are identified as masculine or feminine) implied by THEY being able to stand in for subjects regardless of gender.

The public and polyvocal nature of these Tweets suggests a new interpretation of the traits previously associated with THEY, especially of low individuation or prototypicality. While THEY, in practice, is often completely individuated, the trait of “low individuation” emerges from this data as particular portrayals of #oomf are taken up between seventy-six and 102 accounts. These texts with THEY become more public, the described referent emerging not as one individual, but as many users’ individual #oomf. This, not any feature of the sentence itself, turns the #oomf’s liked to THEY into prototypes. Their low individuation is performed even as multiple accounts further specify the Tweet – their low individuation is the product of the social spread of the message and its application to many specific individuals. Prototypically and low individuation, as found to be characteristics of some THEY Tweets in this corpus, are emergent in the social creation of discourse.

I draw from an ethnographically acquired understanding of Twitter to argue that the intertextual connections in the corpus, rather than individual tokens, provide the most productive analysis of this corpus. This ties my analysis to the corpus studied and questions a common approach to Twitter corpora as “bags of tokens,” unproblematically delimited from each other. Prima facie, the texts of a lightweight social media platform with a simple unit of interaction and very short texts may seem separable, especially when compared to richer modes of interaction. However, in this domain, the interconnection of tokens manifestly produced the most immediately salient difference between them. I hope especially to have shown the usefulness of intertextual connections between Tweets, where much research that

connects Tweets collects data based on social network and analyzes it using such data, which is not as salient to average Twitter users from outside those networks. This intertextual approach also suggests a new way to consider "influence," usually understood as being a matter of popular accounts and social networks, by illustrating how linguistic forms are differently spread within the platform. Previous studies of Tweet dissemination have often focused on content issues with very limited linguistic or social analysis, limiting themselves (albeit in much larger corpora) to computationally-findable features like the presence of URLs and most frequent words. This study asks instead how patterns of dissemination define a linguistic form in discourse.

Finally, the hashtag corpus offers unique possibilities to Twitter researchers. This thesis has related linguistic form to social action, which was exposed in large part due to the approach of collecting a corpus around a hashtag. Though several authors collect data from hashtags (Zappavigna 2011, 2012; Bruns and Burgess 2011) this remains an understudied organizational system in Twitter and is shown here to provide unique and fruitful insight into interconnection of Tweets within my corpus. Perhaps there is something in THEY that allows users to hold it at a distance from themselves, or to find themselves in it and related, but rather than ponder unknowable motivations of users' action, this thesis focuses on the actions taken. By collecting a corpus collected around a hashtag while it was extremely popular, I was able to collect a set of Tweets that show affiliation with each other and that show affiliation with a common referent. Though many studies have used the Twitter API to collect data, the way they query the API leads to very different corpora. Rather than querying by search term, as I did, which collects a thick sample of interrelated Tweets, they make calls to the general Twitter API, which

is often compared to a firehose for its barrage of results, which are notably chaotic and unconnected. Others collect data from specific users or crawl the platform based on user-to-user connections. This hashtag corpus, however, allows for a specific type of social conclusion, tracing not just frequency of Retweets in general, but requiring a bottom-up investigation into the sharing of certain messages. These contexts give insight into how the linguistic content of the Tweet becomes part of discourse.

iii. Limitations, future study

Finally, this study attempts to suggest directions of subsequent research. Future study of THEY may more critically approach the gender binary; future study of Twitter could productively turn to connections emergent between those texts.

Singular THEY is certainly not exhaustively treated here, but I believe the systematic study of THEY in non-generic contexts to be a timely question. My research into singular they with a specific referent is in a large part motivated by my observation of it in my daily life, and the rather mundane corpus collected for this research suggests that that my intuition has a basis in naturally-occurring language. Naturalistic data, as is used here, seems to offer unique insights of where THEY is interchangeable with HE and SHE, while introspection may not produce these as reliably or informatively - as I argue, there is nothing in particular about the sentences studied here that suggests them as THEY-conditioning or SHE- or HE- conditioning sentences.

Also, though it was not treated here, a study of third-person pronouns would ideally admit complication beyond a simple, binary, sex-based identification system of the subjects studied. That is, there have been recent reports of people identifying as “they” (see Language

Log 2013) or using other pronoun neologisms to express self-identification that does not fit neatly into the gender binary. The size of this corpus prevented a Tweet-by-Tweet examination for any other potential non-standard third-person pronouns (especially since no simple or exhaustive list of such possibilities exists). For those #oomf's identified as HE, SHE or THEY the question of gender-identification of the actual referents was far outside the possibilities of the data collected. Still, the availability of THEY as a pronoun of self-identification does suggest a comfort with using it for specific people, and if, indeed, the usage spreads more widely, will certainly affect the possibilities of THEY for all singular people in the future. This is certainly worthy of further examination.

Finally, I believe that there is great potential for the study of intertextual connections in Twitter corpora. While much of the previous research approaches the question of connections between Twitter texts through an interpersonal, network-based analysis, the clusters I study here are found, rather, by intentional connections between the texts themselves, which sometimes fail to encode user-to-user movement. I believe this distinct approach more closely mirrors the life of this discourse among Twitter users; however, in the scale an environment of Twitter, this is still an extremely small study. In order to approach the question of linguistic form, I collected a corpus limited to only one hashtag, imposing a synthetic boundary between Tweets including that string and those not; then, I further delimited my corpus by the inclusion of specific pronouns. These are analytic, not natural, boundaries, and they do not inhere in the text itself or in the emergence of the text. A more complete work will work not only with a larger corpus, but also work through these imposed categories to discover larger connections and trends.

VII. Conclusion

This paper has used taken a bottom-up approach to 'singular they' as linked to the specific antecedent #oomf in a trending topic corpus. Though THEY was previously considered to be unlikely with such an antecedent, it's found to occur in Tweets at close to the same frequency in this corpus as SHE or HE. As no salient differences are found in referent type (as in previous THEY studies), and several examples of similar Tweets crossing between pronouns emerge, it appears that these measures are inadequate to describe the various use of pronouns in this corpus. Instead, it's suggested that this apparent free variation of the pronouns studied in this context should guide us to look outside the sentence in context, much as any approach to the difference between the pronouns HE and SHE would.

For this corpus, at least, there was a clear pattern in the composition of pronoun-including subcorpora: several Tweets including THEY were so widely circulated as to make up a large part of that subcorpus, while all other Tweets in all three subcorpora were disseminated at a much smaller scale, in the vast majority by only one or two users. This reproduction, rather than representing an impoverished set of texts, shows a polyvocality as users share these THEY Tweets, holding the message at a distance from themselves through copying and attribution, and thereby making the Tweets *more* public. As this pattern emerges through social action, it is tied to the situation of texts within the Twitter environment and is therefore inherently tied to the Tweet as a unit of interaction rather than any linguistic form in and of itself. Finally, it is suggested that in this context, though the introspective and token-level analyses of previous studies have not accounted for this data, the features of prototypicality and low individuation discovered by those studies are embodied in social action. Based on the scale that certain texts

that include THEY reach, those texts become *more* public – public not only in the sense of openness for viewing and interaction, but in the actions of users who take up the texts to describe people in their own context. Prototypicality and low individuation emerge as features of THEY in discourse, as the repetition of a text between many users turns the subject into a prototype that stands in for many individuals. Here, meaning is found above the sentence or token level, emergent rather at the level of social action.

Appendix: Annotation of ten random Tweets

Table 1: Features annotated and basis in previous literature.

Feature	Definition	Annotation	Source	Notes
Gender	any indication of referent gender in the text	[+masculine] [+feminine]	--	I do not restrict this to indications of gender that definitely characterize the referent as male or female. Pronouns have been found, in previous studies, to be used in contrast to sex in order to exploit indexicalities (see for example: Rudes and Healy 1979, Mathiot and Roberts 1979, Hall 2002).
Notional Number	number of the referent	[singular] [plural]	Newman 1992, Baranowski 2002	Newman includes a 'neutral' category for sentences that could be either singular or plural; as I argue earlier, #oomf conditions singular referents in this corpus and I therefore take that as the default in ambiguous sentences
Referential Solidity	existence of not of a specific, concrete referent	[+solid] [-solid]	Newman 1992	
Individuation	referent is important based on specific identity rather than type membership	[1-3]	Newman 1998	Newman uses a scale of 1-5. I find this to be untenable in the extremely brief texts of Twitter; instead I make a three-way distinction.

Table 2: Annotation of HE Tweets

Tweet text	Gender	Number	Solidity	Individuation	Notes
RT @SupremeOceans: I don't know why #Oomf is jealous, he's all I tweet about.	--	singular	+solid	3	
"@Trecce_Tree: I wanna chill w/ #oomf this weekend . . but he be playing *sighs*" I can't eem get the whip shawty	--	singular	+solid	3	
#ExplainToMeWhy No matter what I do, or how good I treat #oomf, he still finds it necessary to talk shit about me & treat me like shit	--	singular	+solid	3	
Lmao #oomf aint text back when i asked him a question...i wouldnt of txted back either bro	+masculine	singular	+solid	3	'bro' is a term of address that is derived from 'brother'
I kept texting #oomf knowing that he wasnt gonna reply.....	--	singular	+solid	3	
#oomf better ask me soon or he's gonna be mighty jealous when i go with someone else. #sorrynotsorry	--	singular	+solid	3	
#oomf , said all he do is eat shit & be happy .	--	singular	+solid	3	
RT @puRdie_RHo: "@terriSMASHEM: #oomf look like he can eat the soul out cha. [Unicode smileys]	--	singular	+solid	3	
RT @Teresa_OVO_XO: #Oomf always using them beanies cause he bald..lol	--	singular	+solid	3	
Lmfao #oomf gen write BRB on his TL like he is so special.. U aint gonna b missed	--	singular	+solid	3	[=laughing my fucking ass off, #oomf is going to write 'be right back' on his timeline *[=Tweet it]* like he is so special.. you ain't gonna be missed]

Table 3: Annotation of SHE Tweets

Tweet text	Gender	Number	Solidity	Individuation	Notes
#oomf should come back out here so I can see her face. Then she can leave. Lol	--	singular	+solid	3	
@OhDearOOMF: It hurts to see #oomf with someone else... She doesn't follow me on twitter..	--	singular	+solid	3	
#oomf never made me the video of her whipping her hair back and forth for my birthday... #sadtweet.	--	singular	+solid	3	
RT @BallerBrosman: #oomf didn't even care that I was upset with her today in 10th.	--	singular	+solid	3	
#oomf is thick in a perfect way and stay lookin so damn good she kno she look good lol	+feminine	singular	+solid	3	'thick' is a positive, feminine gendered, physical characterization
smh.. i trying to help #oomf and she don't want help herself...	--	singular	+solid	3	
RT @_SmackThatHoe: #oomf think she a boss! have a _ <3	--	singular	+solid	3	
#Oomf Fake ! I Swear ! I Thought She Was Real AF ! Untill What I Seen In Her Today ! I Think Im Straight On Being Cool With Her After Today[Unicode: 'Victory hand']	--	singular	+solid	3	'AF' abbreviates 'as fuck'
#Oomf don't understand I still got feelings for her.. she stuck on this friend shit	--	singular	+solid	3	
I don't like #oomf because she messed around with #oomf [Unicode: 'Raised hand']	--	singular	+solid	3	

Table 4: Annotation of THEY Tweets

Tweet text	Gender	Number	Solidity	Individuation	Notes
#OOMF need to stop assuming shit fa they fuck sum up. RT !! Chill out wit det b.s you talking .	--	singular	+solid	3	'fa' is a reduced form of 'before.' 'det b.s.' is for 'that bullshit.' This is marked as a RT (presumably of the text before "RT") but not attributed.
Going to polo with #oomf this weekend ! Spending all they money ! \U0001f601\U0001f601\U0001f601\U0001f601	--	singular	+solid	3	
I Was Dwn For A Lil Bit But #oomf Keep Me Up Today...I Really Appreciate Them!	--	singular	+solid	3	"Dwn' abbreviates 'down' as in sad.
#oomf doesn't even know how fake they're being treated by #oomf<<<<	--	singular	+solid	3	
I twatch #OOMF all the time! I'm sure they know already	--	singular	+solid	3	
#oomf Said they wasn't gone text me no more .. Oh !	--	singular	+solid	3	
@thuggchick12: I seen #oomf face pop up on my TL & it made me want to mention them & say bitch be gone .RNS	+feminine	singular	+solid	3	'TL' is for timeline; here, it means that #oomf Tweeted and the Tweet was visible to this user, who follows #oomf
"@OhDearOOMF: If you flirt with #oomf knowing that i like them, i will personally escort you to the gates of hell."	--	singular	+solid	2	
#oomf needs to chill out and realize their life isn't that bad and things could be a hell of a lot worse #bethankful	--	singular	+solid	3	
I'm this close [Unicode 'OK-hand'] in calling #oomf and cussing them the fuck out #PissOff	--	singular	+solid	3	

Bibliography

"#oomf Definition." n.d. Online Dictionary. *Tagdef.com*. http://tagdef.com/oomf.

Balhorn, Mark. 2004. "The Rise of Epicene They." *Journal of English Linguistics* 32 (2): 79–104.

———. 2009. "The Epicene Pronoun in Contemporary Newspaper Prose." *American Speech* 84 (4): 391–413.

Bamman, David, Jacob Eisenstein, and Tyler Schoenebelen. 2012. "Gender in Twitter: Styles, Stances, and Social Networks." http://www.stanford.edu/~tylers/papers.shtml.

Baranowski, Maciej. 2002. "Current Usage of the Epicene Pronoun in Written English." *Journal of Sociolinguistics* 6 (3): 378–397.

Baron, Dennis E. 1986. *Grammar and Gender*. New Haven: Yale University Press.

Baym, Nancy K., and danah boyd. 2012. "Socially Mediated Publicness: An Introduction." *Journal of Broadcasting & Electronic Media* 56 (3) (July 1): 320–329.

Becker, Alton L. 1984. "The Linguistics of Particularity: Interpreting Subordination in a Javanese Text." In *Proceedings of the Annual Meeting of the Berkeley Linguistics Society*, 10:425–436. Berkeley, CA.

Brown, Roger, and Albert Gilman. 1960. "The Pronouns of Power and Solidarity." In *Style in Language*, edited by T. A. Sebeok, 253–276. Cambridge, MA: MIT Press.

Bruns, Axel, and Jean E. Burgess. 2011. "The Use of Twitter Hashtags in the Formation of Ad-Hoc Publics". Conference Paper. *ARC Centre of Excellence for Creative Industries and Innovation; Creative Industries Faculty; Institute for Creative Industries and Innovation*. August 27. http://www.ecprnet.eu/conferences/general_conference/reykjavik/.

Bucholtz, Mary, and Kira Hall. 2005. "Identity and Interaction: A Sociocultural Linguistic Approach." *Discourse Studies* 7 (4-5): 585–614.

Cameron, Deborah. 1992. *Feminism and Linguistic Theory*. New York: St. Martin's Press.

Cha, Meeyoung, Hamed Haddadi, Fabricio Benevenuto, and Krishna P. Gummadi. 2010. "Measuring User Influence in Twitter: The Million Follower Fallacy." In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM)*, 14:8. Washington, DC.

Cheong, M., and V. Lee. 2010. “A Study on Detecting Patterns in Twitter Intra-Topic User and Message Clustering.” In *Proceedings of the 20th International Conference on Pattern Recognition (ICPR)*, 3125–3128. Clayton, Victoria, Australia.

Constant, Noah, Christopher Davis, Christopher Potts, and Florian Schwarz. 2009. “The Pragmatics of Expressive Content: Evidence from Large Corpora.” *Sprache Und Datenverarbeitung* 33 (1-2): 5–21.

Eckert, Penelope. 2008. “Variation and the Indexical Field.” *Journal of Sociolinguistics* 12 (4): 453–476.

———. 2012. “Three Waves of Variation Study: The Emergence of Meaning in the Study of Sociolinguistic Variation.” *Annual Review of Anthropology* 41 (1): 87–100.

Ferrara, Kathleen, and Barbara Bell. 1995. “Sociolinguistic Variation and the Discourse Function of Constructed Dialogue Introducers: The Case of Be + Like.” *American Speech* 70 (3): 265–290.

Gillen, Julia, and Guy Merchant. 2013. “Contact Calls: Twitter as a Dialogic Social and Linguistic Practice.” *Language Sciences* 35 (January): 47–58.

Gruzd, Anatoliy, Barry Wellman, and Yuri Takhteyev. 2011. “Imagining Twitter as an Imagined Community.” *American Behavioral Scientist* 55 (10): 1294–1318.

Hall, Kira. 2002. “‘Unnatural’ Gender in Hindi.” In *Gender Across Langauges: The Linguistic Representation of Women and Men*, edited by Marlis Hellinger and Hadumod Bussman, 133–162. Amsterdam: John Benjamins.

Heim, Irene, and Angelika Kratzer. 1998. “Bound and referential pronouns and elipsis Bound and Referential.” In *Semantics in Generative Grammar*, 239–259. Malden, MA: Blackwell.

Herman, John. 2012. “Introducing #Oomf, Twitter’s Best Ever Hashtag.” *BuzzFeed*. November 21. http://www.buzzfeed.com/jwherrman/introducing-oomf-twitters-best-ever-hashtag.

Herring, Susan C. 2003. “Gender and Power in On-line Communication.” In *The Handbook of Language and Gender*, edited by Janet Holmes and Miriam Meyerhoff, 202–228. Blackwell Handbooks in Linguistics. Malden, MA: Blackwell Publishing.

Holmes, J. 1998. “Generic Pronouns in the Wellington Corpus of Spoken New Zealand English.” *Kotare* 1 (1): 32–40.

Honeycutt, Courtenay, and Susan C. Herring. 2009. “Beyond Microblogging: Conversation and Collaboration via Twitter.” In *Proceedings of the Forty-Second Hawai’i International Conference on System Sciences*. Los Alamitos, CA: IEEE Press. Preprint: http://ella.slis.indiana.edu/~herring/honeycutt.herring.2009.pdf.

Huang, Jeff, Katherine M. Thornton, and Efthimis N. Efthimiadis. 2010. “Conversational Tagging in Twitter.” In *Proceedings of the 21st ACM Conference on Hypertext and Hypermedia*, 173–178. Toronto, Ontario, Canada: ACM.

Java, Akshay, Xiaodan Song, Tim Finin, and Belle Tseng. 2009. “Why We Twitter: An Analysis of a Microblogging Community.” In *Advances in Web Mining and Web Usage Analysis*, edited by Haizheng Zhang, Myra Spiliopoulou, Bamshad Mobasher, C.Lee Giles, Andrew McCallum, Olfa Nasraoui, Jaideep Srivastava, and John Yen, 5439:118–138. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer.

Khosroshahi, Fatemeh. 1989. “Penguins Don’t Care, but Women Do: A Social Identity Analysis of a Whorfian Problem.” *Language in Society* 18 (4): 505–525.

Kiesling, Scott F. 2004. “Dude.” *American Speech* 79 (3): 281–305.

Kozinets, Robert V. 2010. *Netnography: Doing Ethnographic Research Online*. Los Angeles, CA; London: Sage Publications.

Lagunoff, Rachel. 1997. “Singular They”. Doctoral Dissertation, University of California, Los Angeles.

Lavandera, Beatriz R. 1978. “Where Does the Sociolinguistic Variable Stop?” *Language in Society* 7 (2): 171–182.

Leavitt, Alex, Evan Burchard, David Fisher, and Sam Gilbert. 2009. “The Influentials: New Approaches for Analyzing Influence on Twitter”. Pub. 04. Boston: The Web Ecology Project. http://www.webecologyproject.org/wp-content/uploads/2009/09/influence-report-final.pdf.

Liberman, Mark. 2013. “The Future of Singular They”. Blog. *Language Log*. March 8. http://languagelog.ldc.upenn.edu/nll/?p=4482.

Livia, A. 2001. *Pronoun Envy: Literary Uses of Linguistic Gender*. New York: Oxford University Press.

Mackay, Donald G. 1980. “On the Goals, Principles, and Procedures for Prescriptive Grammar: Singular They.” *Language in Society* 9 (3): 349–367.

Marwick, Alice E., and danah boyd. 2011. “I Tweet Honestly, I Tweet Passionately: Twitter Users, Context Collapse, and the Imagined Audience.” *New Media & Society* 13 (1): 114–133.

Mathiot, Madeleine. 1979. “Sex Roles as Revealed Through Referential Gender in American English.” In *Ethnolinguistics: Boas, Sapir and Whorf Revisited*, edited by Madeleine Mathiot, 1–47. The Hague: Mouton.

Matossian, Lou Ann. 1997. “Burglars, Babysitters, and Persons: A Sociolinguistic Study of Generic Pronoun Usage in Philadelphia and Minneapolis”. Philadelphia: University of Pennsylvania.

McConnell-Ginet, Sally. 2011a. “‘What’s in a Name?’: Social Labeling and Gender Practices.” In *Gender, Sexuality and Meaning: Linguistic Practice and Politics*, 169–184. Studies in Language and Gender. New York: Oxford University Press.

———. 2011b. “Prototypes, Pronouns and Persons.” In *Gender, Sexuality and Meaning: Linguistic Practice and Politics*, 185–206. Studies in Language and Gender. New York: Oxford University Press.

Meyers, Miriam Watkins. 1990. “Current Generic Pronoun Usage: An Empirical Study.” *American Speech* 65 (3): 228–237.

Miller, Casey, and Kate Swift. 1976. *Words and Women*. Garden City, NY: Anchor Press.

Mühlhäusler, Peter, and Rom Harré. 1990. *Pronouns and People: The Linguistic Construction of Social and Personal Identity*. Oxford, UK; Cambridge, MA, USA: Basil Blackwell.

Naveed, Nasir, Thomas Gottron, Jérôme Kunegis, and Arifah Che Alhadi. 2011. “Bad News Travel Fast: A Content-Based Analysis of Interestingness on Twitter.” In *Proceedings of the ACM*. Koblenz, Germany.

Newman, Michael. 1992. “Pronominal Disagreements: The Stubborn Problem of Singular Epicene Antecedents.” *Language in Society* 21 (3): 447–475.

———. 1998. “What Can Pronouns Tell Us? A Case Study of English Epicenes.” *Studies in Language* 22 (2): 353–389.

Page, Ruth. 2012. “The Linguistics of Self-Branding and Micro-Celebrity in Twitter: The Role of Hashtags.” *Discourse & Communication* 6 (2): 181–201.

Pauwels, Anne. 2003. “Linguistic Sexism and Feminist Linguistic Activism.” In *The Handbook of Language and Gender*, edited by Janet Holmes and Miriam Meyerhoff, 550–570. Blackwell Handbooks in Linguistics. Malden, MA: Blackwell Publishing.

Pichler, Heike. 2010. “Methods in Discourse Variation Analysis: Reflections on the Way Forward.” *Journal of Sociolinguistics* 14 (5): 581–608.

Potts, Christopher. 2007. “The Expressive Dimension.” *Theoretical Linguistics* 33 (2): 165–198.

Pullum, Geoffrey K. 2010. “Singular They With Personal Name Antecedent”. Blog. *Language Log*. September 1. http://languagelog.ldc.upenn.edu/nll/?p=2600.

Rudes, Blair A, and Bernard Healy. 1979. “Is She for Real? The Concepts of Femaleness and Maleness in the Gay World.” Edited by Madeleine Mathiot. *Ethnolinguistics: Boas, Sapir and Whorf Revisited*: 49–61.

Schiffrin, Deborah. 2006. *In Other Words*. Cambridge: Cambridge University Press.

Semiocast. 2013. “Twitter Reaches Half a Billion Accounts: More Than 140 Millions in the U.S.” *Semiocast*. July 30. http://semiocast.com/publications/2012_07_30_Twitter_reaches_half_a_billion_accounts_140m_in_the_US.

Smith, Aaron. 2011. “Twitter Update 2011”. Main Report. Pew Internet & American Life Project. Washington, DC: Pew Internet. http://www.pewinternet.org/Reports/2011/Twitter-Update-2011/Main-Report.aspx.

Sousa, Daniel, Luís Sarmento, and Eduarda Mendes Rodrigues. 2010. “Characterization of the Twitter @replies Network: Are User Ties Social or Topical?” In *Proceedings of the 2nd International Workshop on Search and Mining User-generated Contents*, 63–70. Toronto, Ontario, Canada: ACM.

Stringer, Jeffrey L, and Robert Hopper. 1998. “Generic He in Conversation?” *Quarterly Journal of Speech* 84 (2) (May 1): 209–221.

Stubbs, Michael. 2001. *Words and Phrases: Corpus Studies of Lexical Semantics*. Oxford; Malden, MA: Blackwell Publishers.

Suh, Bongwon, Lichan Hong, Peter Pirolli, and Ed H. Chi. 2010. “Want to Be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network.” In *Proceedings of the 2010 IEEE Second International Conference on Social Computing (SocialCom)*, 177–184. Minneapolis: IEEE.

Sunden, Jenny. 2002. “‘I’m Still Not Sure She’s a She’: Textual Talk and Types Bodies in Online Interaction.” In *Talking Gender and Sexuality*, edited by Paul McIlvenny, 289–312. Amsterdam: John Benjamins.

Twenge, Jean M, W Keith Campbell, and Brittany Gentile. 2012. “Male and Female Pronoun Use in U.S. Books Reflects Women’s Status, 1900–2008.” *Sex Roles* 67 (9-10) (November 1): 488–493.

Twitter. n.d. “I’m Missing from Search!” *Twitter Help Center*. https://support.twitter.com/groups/32-something-s-not-working/topics/118-search-problems/articles/66018-i-m-missing-from-search#.

———. 2012a. “Using the Twitter Search API.” *Twitter Developers*. August 25. https://dev.twitter.com/docs/using-search.

———. 2012b. “Developer Display Requirements.” *Twitter Developers*. November 20. https://dev.twitter.com/terms/display-requirements.

Weidmann, Urs. 1984. “Anaphoric They for Singular Expressions.” In *Modes of Interpretation: Essays Presented to Ernst Leisi on the Occasion of His 65th Birthday*, edited by Urs Weidman and R Watts, 59–68. Tübingen: Narr.

Wu, Shaomei, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. 2011. “Who Says What to Whom on Twitter.” In *Proceedings of the 20th International Conference on World Wide Web*, 705–714. Hyderabad, India: ACM.

Yang, Zi, Jingyi Guo, Keke Cai, Jie Tang, Juanzi Li, Li Zhang, and Zhong Su. 2010. “Understanding Retweeting Behaviors in Social Networks.” In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, 1633–1636. Toronto, Ontario, Canada: ACM.

Zappavigna, Michele. 2011. “Ambient Affiliation: A Linguistic Perspective on Twitter.” *New Media & Society* 13 (5): 788–806.

Zappavigna, Michele. 2012. *Discourse of Twitter and Social Media*. London; New York: Continuum.

Zwicky, Arnold. 2010. “Singular They Trudges On”. Blog. *Language Log*. January 24. http://languagelog.ldc.upenn.edu/nll/?p=2072.