HAL
open science

“That spelling tho”: A sociolinguistic study of the nonstandard form of though in a corpus of Reddit comments

Marie Flesch

To cite this version:

Marie Flesch. “That spelling tho”: A sociolinguistic study of the nonstandard form of though in a corpus of Reddit comments. European Journal of Applied Linguistics. , 2019, 7 (2), pp.163-188. 10.1515/eujal-2019-0007 . hal-02377670

HAL Id: hal-02377670

https://hal.science/hal-02377670v1

Submitted on 4 Mar 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

"That spelling tho": A sociolinguistic study of the nonstandard form of *though* in a corpus of Reddit comments

Marie Flesch

Abstract
Tho, the nonstandard spelling of *though* which was proposed by American spelling reformers in the 19th century, is making a comeback. In 2013, internet memes such as *that backflip tho* gave a boost to the shortened form. This sociolinguistic study investigates the use of *tho* in RedditGender, a 19 million-word corpus of comments posted by 1044 Reddit users. First, concordance lines generated from the whole corpus were analyzed in order to compare the use of *tho* with the meme and the standard spelling. Then, regression analysis was conducted with a sample of the corpus, containing only complete cases. Results show that *tho* is rarely used in the meme construction that contributed to popularize it, and that it appears more often as an adverb than as a conjunction. They also seem to indicate that the use of *tho* is correlated with gender and race. Most frequent users are black males. This suggests that the shortened spelling is not simply a way to save time when typing, and that it is not semantically equivalent to *though*. It seems to be a marker of affiliation with a social group and of familiarity with internet subcultures.
Keywords: sociolinguistics, internet slang, Reddit, corpus linguistics, nonstandard spellings

Abstract
Tho, die nicht standardisierte Schreibweise, die von amerikanischen Rechtschreibreformern im 19. Jahrhundert vorgeschlagen wurde, feiert ein Comeback. Im Jahre 2013 gaben Internet-Meme wie *that backflip tho* der verkürzten Form einen Schub. Diese soziolinguistische Studie untersucht die Verwendung von *tho* in RedditGender, einem 19 Millionen Wörter umfassenden Korpus von Kommentaren, die von 1044 Reddit-Nutzern veröffentlicht wurden. Sie stützt sich auf eine aus dem gesamten Korpus generierte Konkordanzanalyse einerseits, sowie eine aus einer Probe des Korpus durchgeführte Regressionsanalysen andererseits. Daraus geht hervor, dass *tho* selten in der Meme-konstruktion verwendet wird, die dazu beigetragen hat, es zu popularisieren, und dass es häufiger als Adverb denn als Konjunktion erscheint. Die soziolinguistische Untersuchung scheint auch darauf hinzudeuten, dass die Verwendung von *tho* mit Geschlecht und Rasse korreliert ist. Die häufigsten Nutzer sind schwarze Männer. Dies deutet darauf hin, dass die verkürzte Schreibweise nicht nur eine Möglichkeit ist, beim Tippen Zeit zu sparen, und dass sie semantisch nicht gleichbedeutend ist mit *though*. Es scheint ein Zeichen der Zugehörigkeit zu einer sozialen Gruppe und der Vertrautheit mit Internet-Subkulturen zu sein.
Keywords: Soziolinguistik, Internet Slang, Reddit, Korpuslinguistik, Memes

Abstract
Tho, la ortografía atípica de *though* propuesta por los reformadores estadounidenses en el siglo XIX, está volviendo a aparecer. En 2013, los memes de Internet como *that backflip tho* dieron un impulso a la forma acortada. Este estudio sociolingüístico investiga el uso de *tho* en RedditGender, un corpus de 19 millones de palabras de comentarios publicados por 1044 usuarios de Reddit. Al principio, se analizaron las líneas de concordancia generadas a partir de todo el corpus para comparar el uso de *tho* con el meme y la ortografía estándar. Luego, se realizó análisis de regresión con una muestra del corpus que contenía sólo casos completos. Los resultados muestran que *tho* es poco utilizado en la construcción del meme que contribuyó a popularizarlo, y que aparece más a menudo como un adverbio que como una conjunción. También parecen indicar que el uso de *tho* está correlacionado con el género, y la raza. Los usuarios más frecuentes son hombres negros. Esto sugiere que la ortografía acortada no es simplemente una forma de ahorrar tiempo al escribir, y que no és semánticamente equivalente al *though*. Parece indicar afiliación a un grupo social y marcar familiaridad con las subculturas de Internet.
Keywords: Sociolingüística, internet slang, Reddit, lingüística de corpus

Abstract
Tho, la graphie alternative de *though* proposée par les réformateurs de l'orthographe américains au 19ème siècle, fait un retour remarqué. Depuis 2013, elle connait une nouvelle vie sur internet grâce au succès de mèmes comme *that backflip tho.* Cette étude sociolinguistique examine l'utilisation de *tho* dans RedditGender, un corpus de 19 millions de mots composé de commentaires mis en ligne par 1044 utilisateurs sur le site américain Reddit. Elle s'appuie à la fois sur l'analyse de concordances, réalisée sur l'intégralité du corpus, et sur des modèles de régression créés à partir d'un échantillon du corpus. Les résultats montrent que *tho* est rarement utilisé dans la construction popularisée par les mèmes, et que la graphie semble être plus facilement adoptée comme adverbe que comme conjonction. L'étude sociolinguistique révèle que l'utilisation de *tho* semble être corrélée avec le genre et la race des Redditeurs. Les internautes qui utilisent *tho* le plus fréquemment sont des hommes africains-américains. Ces résultats suggèrent que *tho* n'est pas simplement une graphie raccourcie de *though*. *Tho* semble être un marqueur d'appartenance à un groupe social et le signe d'une familiarité avec les sous-cultures d'internet.
Mots-clés : sociolinguistique, Reddit, linguistique de corpus, orthographe

1 Introduction

Since the 1990s, lexical features of CMC have been an object of research for sociolinguists. Scholars have studied communication styles of men and women (Herring 1994, 2003; Herring & Paolillo 2006), and emoticons and emoji use (Witmer & Katzman 1997; Baron 2004; Lee 2003; Prada et al. 2018). Nonstandard spellings, however, seem not to have drawn the attention of researchers as much as other CMC

features. Nonetheless, researchers who have investigated phenomena such as expressive lengthenings and the omission of apostrophes (Coats 2017; Squires 2012) have shed a light on how nonstandard spellings can contribute to create social meaning through their correlation with gender. A lot still remains to be uncovered, online and offline, as Sebba points out: "the study of the nuts-and-bolts of writing - the scripts and orthographies which form the basic medium for written expression - has been largely neglected from a social point of view" (2012: 1). The internet provides the perfect space to explore this, as the written word has never been as accessible to researchers. This paper sets out to explore the topic of nonstandard spellings by focusing on a single word: *tho*, the alternative form of *though*. It looks at the way *tho* is used on Reddit, a popular American community website, and presents a quantitative study of a 19 million-word corpus of Reddit comments, drawing on demographic data gathered from the content posted by 1044 Redditors. It starts by providing a review of the literature about CMC and gender, age and race, and an overview of the Reddit website. It then goes on to retrace the history of the nonstandard spelling of *though*, before describing the RedditGender corpus and the quantitative analyses performed. Finally, it presents and discusses the results, and points outs the limitation of the study.

2 Background

2.1 Gender and CMC

Gender has been extensively studied by sociolinguists (Labov 1972; Trudgill 1974; Eckert & McConnell-Ginet 1995; Schilling-Estes 2002). In CMC research, it has been an object of interest since the 1990s. Early studies of gender and CMC uncovered differences in communicative styles of men and women. Herring (1994, 2003) described a female style characterized by supportiveness and attenuation, and an adversarial male style. Thomson and Murachver (2001) identified gender differences in many features, including a higher frequency of questions, intensive adverbs and modals in females. Gendered communicative styles were also explored by Colley and Todd (2002), who found that women used more questions and more exclamation marks.

One of the most widely investigated CMC features, in regards to gender, is emoticon use. A majority of studies has reported that women tend to use more emoticons than men. It was shown by corpus studies of newsgroups (Witmer & Katzman 1997), Instant Messaging (Baron 2004; Lee 2003), text messaging (Tossell et al. 2012), chatrooms, (Del Teso-Craviotto 2008), Facebook (Oleszkiewicz et al. 2017), and Twitter (Coats 2017). Emoji use seems to follow the same pattern (Prada et al. 2018). However, some studies have found no difference in emoticon use between men and women (Luor et al. 2010; Thompson & Filik 2016; Ogletree, Fancher & Gill 2014), or have found that men used more emoticons than women (Huffaker & Calvert 2005). Studies of gender and nonstandard orthography have also uncovered both differences and similarities between genders. Baron (2004) found that women were

more likely to use full words than males. By contrast, Cougnon and François (2010), in their study of Belgian French text messages, noted that women's messages tended to be more abbreviated than men's. Herring and Zelenkauskaite, in their (2008) study of Italian text messages, came to the same conclusion. They also reported that women used more expressive punctuation. Use of exclamation points was found to be linked with the female gender by Waseleski (2006). In a corpus of Flemish Dutch chatspeak, Peersman et al. (2016) did not find any effect of gender on the likelihood of producing abbreviations, acronyms, and expressive lengthenings.
Several other studies link expressive lengthenings such as *fuuuck*, *nooo*, and *aaaahhhh* to females (Rao et al. 2010, Bamman, Eisenstein & Schnoebelen 2014, Coats 2017). Squires (2012), who studied the omission of apostrophes in contractions and possessives, found that males were less likely to use apostrophes than females.
Very few, if any, studies of CMC and language have considered gender outside of the cisgender male/cisgender female binary. The language of non-binary people, who do not place themselves in the binary gender system and identify as bigender, genderfluid, genderqueer or agender, is still to be explored. Gratton (2016), in her pioneering study of non-binary speech, calls for more linguistic research on non-binary individuals, as it could also help to shed more light on the ways females and males use language.

2. 2 Age and CMC

Even though age has been less of a focus of CMC research than gender, a number of studies have explored the way it impacts language use on the internet. Emoticon use, for instance, has been found to be negatively correlated with age in corpus studies (Oleszkiewicz et al. 2017; Sánchez-Moya & Cruz-Moya 2015). In studies based on surveys, younger participants have reported using more emoji and emoticons (Settanni & Marengo 2015; Prada et al. 2018). In their study of Whatsapp, Sánchez-Moya & Cruz-Moya (2015) noted that teenagers used more stylized spellings, such as *veeery*, than adults. Nonstandard spellings and abbreviations were found to be features more readily adopted by younger users (Cougnon & François 2010). Furthermore, teenagers and young adults tend to write shorter posts than older users (Finlay 2014). Finlay also points out the dearth in quantitative studies of language and age in CMC, likely due to the difficulty of obtaining large datasets containing age information.

2. 3 Race and CMC

The sociolinguistic study of race emerged in the 1960s in the US, with the description of African American Vernacular English (Labov 1972; Smitherman 1977; Baugh 1983), Chicano English (Fought 2002) and Native American English (Leap 1993). Other scholars have studied how racial identities are constructed through language (Bailey 1996; Zentella 1997; Cutler 1999; Bucholtz 1999). Researchers have also explored how race is constructed online (Kolko, Nakamura & Rodman 2000), and geek identity as a racialized (white) category (Varma 2007; Kendall 2011). From a

sociolinguistic perspective, the cyberspace remains quite unexplored. Bailey (1996) argues that Netiquette and acronyms constitute "an unwelcoming terrain for marginalized cultures" (p. 22) and that geeks, the primary occupants of the cyberspace, use language to keep outsiders, meaning women, Hispanics and blacks, at bay. Asian Americans benefit from a "cultural neutrality" and are not excluded from geek spaces (Bailey 1996: 22). More research is certainly needed to understand how race is expressed through language in CMC, and to what extent African American Vernacular English influences the "cool" subcultures of the internet.

2. 4 Reddit

Reddit is an American community website where users can submit links and post original content. They can also create their own communities, called "subreddits". The names of subreddits are always preceded by the prefix r/, as in r/politics or r/relationships. Each subreddit is moderated independently by volunteers, who set and enforce the rules. Users decide which posts and comments are the most relevant and interesting through a system of upvoting/downvoting. Upvotes and downvotes also determine the "karma" of a Redditor, i.e. their popularity on the site. It is not necessary to be a member of the site to access subreddits and view threads, but it is in order to post content. Accounts are pseudonymous and easy to set up, and Redditors can create and use several profiles. By contrast with social media sites such as Facebook, Twitter or Instagram, Reddit profiles are basic, and do not contain pictures or descriptions.

Since its creation in 2005, Reddit has become "a popular center of Geek culture". It reflects geek interests such as computing, gaming, and science, and often displays misogynistic and racist ideas (Massanari 2017: 311). Most of the user base is male, with only 33% of US Reddit users being female (Barthel et al. 2016). According to Massanari, the site has been built in a way that makes "participation difficult for women and people of color" (p. 332). Reddit is thus an interesting platform for researchers who want to explore gender and race in the cyberspace, especially since interactions are often open and candid because of their pseudonymous nature.

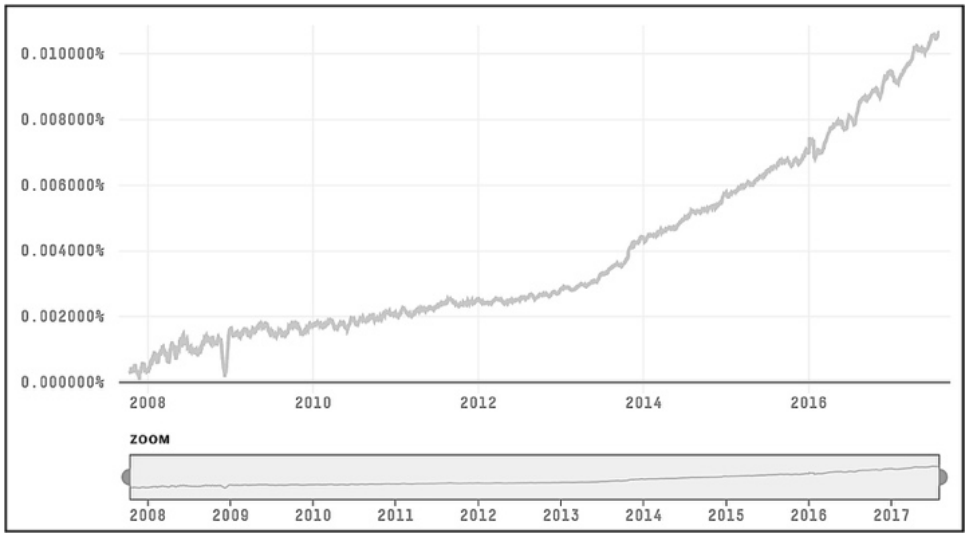
2. 5 A history of *tho*

In the late 19th century, American spelling reformers advocated for the use of shortened forms such as *tho*, *thru*, *catalog*, *gard*, *giv*, or *liv* (Marshall 2011). The shortened spellings of *though* and *through* were again proposed by the Simplified Spelling Board in 1906 (Ranow 1954) but they seem to have never really caught on ("Tho", n.d.). Today, however, *tho* is making a remarkable comeback online as one of the nonstandard spellings which, together with acronyms, emoticons and abbreviations, make up "Internet slang".

Tho came to the attention of linguists in 2010, when it was found to be one of the most frequent nonstandard elements in Kemp's corpus of "textisms" (2010). It was also one of the CMC forms Crystal found in his corpus of tweets (2012), and in Tagliamonte's corpus of email, instant messaging and texting (2016).

The Reddit Ngram viewer (King & Olson 2015), which allows to search Reddit comments from late 2007 to July 2017, is helpful to understand the rise of *tho*. It shows a steep and steady increase in the use of *tho* on the American community website (Figure 1). The shortened form seems to have picked up momentum around late 2013, at a time when several *tho* memes circulated on the web. They often followed the construction *that [noun] tho*, sometimes adopting the alternate spelling *dat [noun] doe*. In this construction, *tho* is used "to place a positive emphasis on a particular aspect or feature within a story, image or video that has been shared online" according to the website Know Your Meme ("Dat Tho", n.d.).
This same site tried to retrace the history of the meme; it suggests that the slang expression *dat ass*, which was posted on the imageboard website 4chan around 2009, was a precursor of *that [noun] tho*. The meme appears to have spread with the video "Dat Dagger Tho" by gaming YouTuber TSirDiesAlot, which was posted in April 2013, and most notably with a video posted by KingBach in June 2013 on the defunct video service Vine, which received 620,000 likes in a year. Captioned "#ButThatBackflipTho", it shows a young man doing a backflip instead of chasing the thief who has just stolen a woman's purse (Figure 2). The man then proudly announces to the camera "Yeah but that backflip though!", pointing out that even though he did not help the woman, his backflip was still impressive.

Figure 1: Result of a search for *tho* on the Reddit Ngram viewer



The *Urban Dictionary* seems to feature one of the oldest definition of *tho* in its CMC use. It dates back to 2005 and says: "Short for 'though'. " Most often found in IRCs, IMs, and blogs." (Urban Dictionary: tho, n.d.). Even if not considered a reliable source, the *Urban Dictionary* gives more information about this new use of *tho*, with three other definitions. Its 2009 definition, provided by an internet user named E.Z.E, specifies that "Tho. is the perfect period. It can go after any comment. Tho. works every time. Try it." The next definition, in a chronological order, links *tho* to hip-hop

culture, with a reference to two rappers and songwriters: "Slang word for "though" and used far too often by people who ride fixie bikes, hypebeasts, and any other person(s) who associates themselves with the media. More or less to followers of Drake and Kid Cudi." The more recent definition, dated April 6, 2015, reaffirms the "punctuating" function of *tho*: " An overused form of the word 'though'. Usually placed at the end of sentences" (Urban Dictionary: tho, n.d.).
The *y tho* meme, which was posted on Imgur in December 2014, also probably contributed to the spread of *tho* ("Y tho", n.d.). Associated with a painting of Pope Leon X by Botero (Figure 3), it is, according to the Know Your Meme site, "a popular slang phrase usually asked in a trolling manner in response to a senseless action or statement".

Figure 2: *That backflip tho* meme[1]

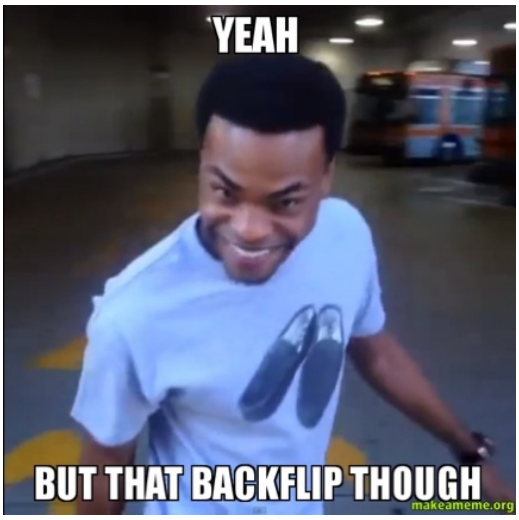
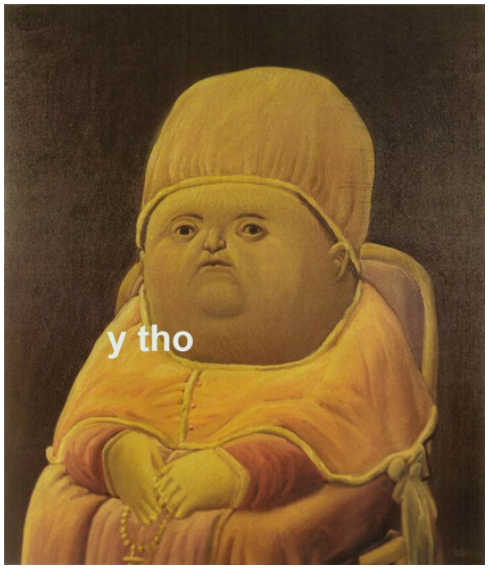


Figure 3: *y tho* meme



3 Methods

3.1. RedditGender

The RedditGender corpus was built by the author of the study as part of her PhD research, which is ongoing, and will be made available to researchers after the author's PhD defense. It is made up of comments written by 1044 Redditors, and contains around 19 million tokens. The corpus was designed with the purpose of conducting a sociolinguistic study of CMC and gender; random sampling was excluded, Reddit being an anonymous platform. Convenience sampling was used instead; most Redditors were selected in forums dealing with gender issues, such as r/AskWomen, r/AskMen or r/transgender, since gender was the primary focus of corpus construction.

[1] The caption was added on a screenshot by an anonymous internet user. KingBach captioned his video #ButThatBackFlipTho

Redditors were selected based on several criteria. First, the quantity and length of comments posted were taken into account. Since the corpus was built for the study of CMC features, some of which are somewhat rare, each subcorpus needed to be quite large. It was also necessary for the subcorpora to be balanced, in order to be able to compare relative frequencies of various lexical elements. Most subcorpora contain between 16,000 and 18,000 tokens.
Second, only users who clearly indicated demographic information about the gender they identify with were chosen. Each profile was manually inspected to verify if it contained other demographical data. Two means of obtaining information were used. First, searches were performed in the Redditors' comment history by using the Ctrl + F search functionality on the Chrome browser. Different keywords were used, such as "I'm a", "I live in" or "I was born in". This investigative strategy proved to be successful in many cases, as is shown in Table 1.

Table 1. Examples of search results using the keywords "I'm a"

I'm a black man in R & D and this doesn't happen to me.
I'm a transwoman who used to sing Tenor I parts and thinks it would be nice to raise my range.
I'm a Hispanic too and I let my mom know that I wasn't planning having kids.
I'm a Chinese-Canadian and I hate the double standards that are going on here.
I'm a 40s hetero male in middle management.
I'm a 46 year old boring accountant lady who likes kittens.
I'm a 31 year old trans woman.
I'm a 42-yr-old female from the Pacific Northwest.
I'm a young 20s black woman in Washington (formerly California).
I'm a lesbian living in Kansas, and whatever security I felt in my future before was gone in an instant.
I'm a middle aged guy, military veteran, pretty conservative.
I wanna be involved!! I'm a nonbinary trans woman of color!!
I'm a white girl and I love Starbucks.

Demographic information was also found in the Redditors' "flairs". On Reddit, "flairs" are little boxes which contain information usually related to the topic of subreddits. They are optional and subreddit-specific. For instance, on r/AskMenOver30 and r/AskWomenOver30, flairs indicate the age range a person belongs to. On r/asktransger, flairs provide information about gender, sexual orientation, and age. Gender was the main variable needed, but in many cases age, sexual orientation, occupation, race and country were also collected. Since Reddit is predominantly white, heterosexual, and male (Barthel et al., 2016), it was decided to over-represent certain categories of the site user base, such as female, LGBTQ, Hispanic, Asian and black users, in order to be able to study the interaction of gender with other variables.

The comments were collected between March 2017 and July 2017, but often date back to several weeks to two years prior. Of the 1044 Redditors included in the corpus, 78% are American; 372 are male, 372 are female, 100 are trans-women, 100 are trans-men and 100 are non-binary. It was not possible to collect data about race for all Redditors. Only 460 users gave information about their race background, of which 92 are black, 68 are Hispanic, 203 are white, 69 are Asian, and 50 are “Other” (Arab, Native American, mixed race, etc.). “Hispanic” is not considered a race in the US Census, but it was included in the race category on the ground of a Pew Research Center survey, which found that two-thirds of Hispanic people consider their Hispanic background as part of their racial background (Gonzalez-Barrera & Lopez 2015).

The exact age (in number of years) was collected for 820 Redditors. In some cases, only an approximation of the age of the Redditor was obtained. It was mostly the case when "flairs" indicating age were used as a source of information. On r/AskWomenOver30 and r/AskMenOver30, these flairs usually indicate an age range, such as "30-34", or "25-29". Because of this, and in order to increase the number of complete cases, it was decided to use age categories instead of age as numerical variable. Five categories were created: 1 (14-19 years old), 2 (20-29 years old), 3 (30-39 years old), 4 (40-49 years old) and 5 (50 years old and older).

3.2. Statistical analyses

The corpus was encoded into XML-TEI TXM, a format designed to work with the free and open-source program TXM. Concordances for *tho* were generated in TXM and then analyzed with R. First, a study of the frequency of *tho* was conducted in the whole corpus, which contains the contributions of 1044 individuals (Table 2). Then, a sociolinguistic analysis was performed, and regression models were fitted. This analysis does not explore the corpus in its entirety, because obtaining demographic information about Redditors was not always fruitful. When retaining gender, age, and race as predictor variables, the dataset contains 576 missing observations. A second dataset containing 474 complete cases was thus created for the sociolinguistic analysis (Table 3) and was used to fit the first regression model.

Table 2. RedditGender: composition of the corpus

Variables and their levels	Redditors
Gender	
Cisgender females	372
Cisgender males	372
Transgender females	100
Transgender males	100
Non-binary individuals	100
Age	
14-19	106
20-29	503
30-39	300
40-49	91
50-59	30
Unknown	14
Race	
White	203
Black	92
Asian	69
Hispanic	68
Other	50
Unknown	562
Total	**1044**

Table 3. Dataset used to fit the first regression model

Variables and their levels	Redditors
Gender	
Cisgender females	194
Cisgender males	191
Transgender females	36
Transgender males	27
Non-binary individuals	26
Age	
14-19	50
20-29	261
30-39	121
40-49	31
50-59	11
Race	
White	201
Black	91
Asian	69
Hispanic	65
Other	48
Total	**474**

The dataset contains a lot of zero values for the dependent variable, i.e. the frequency of *tho* in each subcorpus. Since only 151 Redditors out of 474 used *tho*, there are 323 zero counts. This extremely skewed distribution made it impossible to perform linear regression or Poisson regression. Instead, a zero-inflated count model was created with the `zeroinfl` function from the `pscl` package created by Zeilis et al. (2008). The choice was made to use a zero-inflated negative binomial model instead of a zero-inflated Poisson model. The two models were fitted and compared with the boundary likelihood ratio test. The test showed that the zero-inflated negative binomial model fits better than the zero-inflated Poisson model, because it is better at dealing with overdispersed data (Hilbe 2014).
The `zeroinfl` function returns a two-component mixture model. Its first part shows the probability of outcomes greater than zero. The second component is a binomial model which describes the probability of a zero count (Faraway 2016). Here, it describes the probability of a Redditor not having *tho* in their linguistic repertoire.
A second zero-inflated negative binomial regression model was created in order to explore interactions between race and gender, with a reduced dataset (Table 4). It only takes into account cisgender Redditors, since some interactions with trans Redditors were nonexistent. In order to explore into more detail and nuance the relationships between age, race and gender, two other zero-inflated negative binomial models were created, the first with cisgender women only, and the other with cisgender men only. The procedure was not repeated with the trans and non-binary groups, because of the small size of samples.

Table 4. Dataset used to fit the second regression model with interactions

Variables and their levels	Redditors
Gender	
Cisgender females	194
Cisgender males	191
Age	
14-19	39
20-29	205
30-39	103
40-49	27
50-59	11
Race	
White	127
Black	88
Asian	63
Hispanic	64
Other	43
Total	**385**

4 Results

4.1 Use of *tho* in the corpus

Tho appears 1205 times in the corpus (0.006% of tokens), while the conventional spelling *though* occurs 17,709 times (0.09% of tokens). As a comparison, *thru*, the shortened form of *through*, appears only 181 times (0.0009%), with *through* having a frequency of 11,396 (0.06%). The standard spelling of *though* is only 14.7 times more common than *tho*, while *through* is 62.9 times more frequent than *thru*. *Doe*, a variant associated with African-American Vernacular English (McCulloch 2015) was much rarer than *tho*; it appeared only 46 times (0.0002%). *Tho* was used by 296 Redditors (28.4%) in the RedditGender corpus.

Odds ratios were computed for each Redditor, in order to see if some of them had completely replaced *though* by *tho*. In the whole dataset, the use of *though* compared to *tho* is inferior to 1 in 46 cases, meaning that these Redditors used *tho* more often than *though*. In 10 cases, the odds ratio is equal to zero; these Redditors did not use *though* at all, to the benefit of *tho*. Five Redditors have an odds ratio equal to 1, and used both variants to the same extent.

Analysis of the concordance lines showed that *tho* does not often occur in the meme construction *that [noun] tho*. Even when all the variants of the meme were taken into account (Table 5), the structure appeared only 53 times. The meme structure also occurred twice with the traditional spelling of *though*. The *y tho* meme was a lot less frequent than *that [noun] tho*: it was used only 3 times. Further inspection of the concordance lines was conducted in order to see if the shortened spelling had spread to the conjunction *though*, or if it was only used as an adverb, as in the meme. The *Oxford Living Dictionaries tho* entry certainly suggests that it is possible; 9 of its 16 example sentences use *tho* as a conjunction ("tho'", n.d.), even though it is unclear where these examples come from. In the corpus, however, *tho* is overwhelmingly used as an adverb, with a frequency of 1016. Only 81 *tho* tokens were conjunctions. By contrast, in a random sample of 1000 standard spellings of *though*, almost a third (306) were used as conjunctions. *Doe* was always used as an adverb.

Table 5. Meme constructions of *tho* used in the corpus

Variant	Examples from the corpus	Raw frequencies
dat [blank] tho	DAT ASS THO But dat cute suit tho.	15 (27.3% of meme contructions)
dem [blank] tho (plural form)	Dem arms tho. But dat beat tho.	6 (10.9%)
that [blank] tho	That guitar riff tho, and the ending is so cool : (That last pic tho!! Kill it girl	13 (23.6%)
the [blank] tho	B-b-but the castration tho But the glow on the bride tho! Mixed babies ftw.	5 (9.1%)
Others	Damn son, those eyes tho! Cats tho < 3	14 (25.5%)
that [blank] though	Wow, that cast though. But that achievement though	2 (3.6%)

4.2 Sociolinguistic analysis

4.2.1 Descriptive statistics

Descriptive statistics performed on the whole corpus (Tables 6 and 7) seem to show trends in the use of *tho*, and particularly a possible correlation between race, age, and the frequency of the shortened form. Use of *tho* seems to decrease with age. On average, black and Hispanic Redditors produced respectively 0.13 and 0.15 nonstandard spellings of *though* per 1000 words, while Asian and white Redditors used *tho* only 0.03 and 0.05 time per 1000 words. Even though Redditors identifying as blacks and Hispanics only make up 15.4% of the sample, they produced 33.4% of all the nonstandard spellings of *though* (N=403).

Table 6. Use of *tho* in the whole corpus, per ethnic group

	Redditors who used *tho* at least once	Frequency of *tho*, per 1000 words
Black Redditors (N = 92)	38 (41.3%)	0.13
Hispanic Redditors (N = 68)	30 (44.1%)	0.15
White Redditors (N = 202)	56 (27.6%)	0.05
Asian Redditors (N = 69)	13 (18.8%)	0.03
Others (N = 50)	18 (36%)	0.05
All Redditors (N = 1044)	296 (28.4%)	0.06

Table 7. Use of *tho* in the whole corpus, per age group

Age groups	Reditors who used *tho* at least once	Frequency of *tho*, per 1000 words
14-19 (N = 106)	53 (50%)	0.13
20-29 (N = 503)	158 (31.4%)	0.07
30-39 (N = 300)	61 (20.3%)	0.04
40-49 (N = 91)	14 (15.4%)	0.04
50 + (N = 30)	4 (13.3%)	0.02
Age unknown (N = 14)	6 (42.8%)	0.06
All Redditors (N = 1044)	296 (28.4%)	0.06

4.2.2 Zero-inflated binomial regression

The binomial model computed with the smaller sample, containing only complete cases, describes the probability of *tho* not being used at all (Table 8). It only yields significant results for the race category. It reveals that in the black group, the odds of not using *tho* at all are reduced by a factor of 0.45 ($p < 0.05$). In other words, black Redditors are the most likely to use *tho*. The other component of the model did not produce any significant result as to the effect of gender, age and race on the frequency of use of *tho*.

4.2.3 Zero-inflated binomial regression with interactions

The count part of the model with interactions shows two significant interactions of race and gender. Asian-American and black cisgender males used *tho* significantly more than white females, the reference level of the model. However, it is to be noted that there is one outlier in the small number of cisgender male Asian Redditors present in the sample. Out of 30 male Asian Redditors, only 6 used *tho* at all, the outlier using it at a frequency of 0.83 times per 1000 tokens. The fact that male black Redditors use *tho* at a rate 3.31 times higher than that of white females seems more significant, 19 of 37 black males using *tho* at least once, with a mean of 0.21 *tho* per 1000 words.

Table 8. Output of zero-inflated negative binomial regression, with coefficients, exponentiated coefficients, standard errors, z-values and p-values

Count model coefficients (negative binomial with log link):

	Estimate	Exp. Coeff	Std. Error	z-value	Pr(>|z|)
(Intercept)	3.00854	20.26	0.25410	11.840	<2e-16
gender:cis-male	0.21146	1.23	0.22490	0.940	0.347
gender:trans-female	-0.15090	0.86	0.40602	-0.372	0.710
gender:trans-male	-0.15806	0.85	0.41862	-0.378	0.706
gender:non-binary	0.01411	1.01	0.44131	0.032	0.974
age group:2	0.23002	1.26	0.56363	0.408	0.683
age group:3	-0.20973	0.81	0.47149	-0.445	0.656
age group:4	-0.32451	0.72	0.43862	-0.740	0.459
age group:5	-0.54013	0.58	0.33206	-1.627	0.104
race:black	0.33981	1.40	0.28445	1.195	0.232
race:asian-am	-0.45483	0.63	0.37664	-1.208	0.227
race:hispanic	0.49585	1.64	0.31215	1.589	0.112
race:other	-0.49625	0.61	0.34667	-1.431	0.152

Zero-inflation model coefficients (binomial with logit link):

	Estimate	Exp. Coeff	Std. Error	z-value	Pr(>|z|)
(Intercept)	1.4934	4.45	0.2952	5.058	4.23e-07
gender:cis-male	-0.4435	0.64	0.2436	-1.820	0.0687
gender:trans-female	-0.7430	0.47	0.4530	-1.640	0.1009
gender:trans-male	-0.4121	0.66	0.4888	-0.843	0.3992
gender:non-binary	-0.6562	0.52	0.4984	-1.317	0.1880
age group:2	1.0791	2.94	0.5825	1.852	0.0640
age group:3	-0.3721	0.69	0.5024	-0.741	0.4589
age group:4	-0.3972	0.67	0.4484	-0.886	0.3757
age group:5	-0.1360	0.87	0.3296	-0.413	0.6799
race:black	-0.7953	0.45	0.3136	-2.536	0.0112
race:asian-am	0.4134	1.51	0.3808	1.085	0.2777
race:hispanic	-0.5773	0.56	0.3446	-1.675	0.0939
race:other	-0.5837	0.56	0.3929	-1.486	0.1373

Table 9. Output of zero-inflated negative binomial regression with interactions, with coefficients, exponentiated coefficients, standard errors, z-values and p-values

Count model coefficients (negative binomial with log link):

	Estimate	Exp. Coeff	Std. Error	z-value	Pr(>|z|)
(Intercept)	3.32521	27.80	0.33869	9.818	<2e-16
gender:cis-male	-0.47195	0.62	0.50930	-0.927	0.3541
race:black	-0.34332	0.71	0.47583	-0.722	0.4706
race:asian-am	-1.33082	0.26	0.60383	-2.204	0.0275
race:hispanic	0.43255	1.54	0.53586	0.807	0.4196
race:other	-0.69843	0.50	0.52257	-1.337	0.1814
age group:2	-0.13857	0.87	0.60046	-0.231	0.8175
age group:3	-0.19913	0.82	0.47523	-0.419	0.6752
age group:4	-0.34216	0.71	0.42867	-0.798	0.4248
age group:5	-0.33964	0.71	0.35487	-0.957	0.3385
cis-male:black	1.19903	3.32	0.58382	2.054	0.0400
cis-male: asian-am	1.60373	4.97	0.77768	2.062	0.0392
cis-male:hispanic	0.15395	1.17	0.65123	0.236	0.8131
cis-male:other	0.28959	1.34	0.75129	0.385	0.6999

Zero-inflation model coefficients (binomial with logit link):

	Estimate	Exp. Coeff	Std. Error	z value	Pr(>|z|)
(Intercept)	1.55766	4.75	0.38662	4.029	5.6e-05
gender:cis-male	-0.42029	0.66	0.46574	-0.902	0.3668
race:black	-0.92074	0.40	0.47469	-1.940	0.0524
race:asian-am	-0.12576	0.88	0.60461	-0.208	0.8352
race:hispanic	-0.40827	0.66	0.54832	-0.745	0.4565
race:other	-1.13190	0.32	0.55837	-2.027	0.0426
age group:2	0.94399	2.57	0.59490	1.587	0.1126
age group:3	-0.55112	0.58	0.52166	-1.056	0.2908
age group:4	-0.20311	0.82	0.46171	-0.440	0.6600
age group:5	0.08282	1.09	0.34785	0.238	0.8118
cis-male:black	-0.22383	0.80	0.65510	-0.342	0.7326
cis-male: asian-am	0.53613	1.70	0.81142	0.661	0.5088
cis-male:hispanic	-0.67229	0.51	0.71592	-0.939	0.3477
cis-male:other	0.64915	1.91	0.84655	0.767	0.4432

4.2.4 Zero-inflated binomial regression models, gender-specific

Modelling the use of *tho* by cisgender males and females separately also shed some light on the interaction of gender and race. Two zero-inflated negative binomial models were fitted with a subset of the dataset used in previous regression models, with only cisgender males (N=191) and cisgender females (N=194) (Table 10).
The probability of using *tho* at all was the strongest for black males and females, and Hispanic males ($p < 0.05$). Among male Redditors who used *tho*, black males were the most frequent users of the nonstandard spelling. No such effect was observed in the female group. Asian American females who have *tho* in their repertoire were those who used it the least ($p < 0.05$), producing it at a rate 0.29 times lower than that of white women. Age also had a significant effect in the female group, with females aged 50 and older being the least frequent users of *tho.* Their production of *tho* is 0.64 times lower than that of the youngest females ($p < 0.05$).

Table 10. Output of zero-inflated negative binomial regression for the male and the female groups, with coefficients, exponentiated coefficients, and p-values

Cisgender males				**Cisgender females**			
Count model coefficients							
	Estimate	Exp. Coeff	Pr(>|z|)		Estimate	Exp. Coeff	Pr(>|z|)
(Intercept)	2.31761	10.15	4.75e-09	(Intercept)	3.16789	23.76	<2e-16
age group:2	-1.37651	0.25	0.11199	age group:2	0.56419	1.76	0.4619
age group:3	-0.26326	0.77	0.73222	age group:3	0.17430	1.19	0.8059
age group:4	0.15464	1.17	0.85186	age group:4	-0.20809	0.81	0.6920
age group:5	0.43712	1.55	0.43764	age group:5	-1.05527	0.35	0.0440
black	0.94037	2.56	0.00897	black	-0.23312	0.79	0.6266
asian-am	0.38121	1.46	0.44283	asian-am	-1.20817	0.30	0.0408
hispanic	0.68319	1.98	0.07263	hispanic	0.58369	1.79	0.2596
other	-0.21528	0.81	0.69472	other	-0.47691	0.62	0.3647
Zero-inflation model coefficients							
(Intercept)	1.2835	3.61	0.000942	(Intercept)	1.4652	4.33	0.000445
age group:2	1.0233	2.78	0.242474	age group:2	1.1066	3.02	0.216613
age group:3	-0.2685	0.76	0.724339	age group:3	-1.4726	0.23	0.067802
age group:4	-0.8077	0.45	0.316210	age group:4	0.4128	1.51	0.505425
age group:5	-0.5659	0.57	0.336263	age group:5	0.9180	2.50	0.093306
black	-0.9625	0.38	0.036558	black	-1.1242	0.32	0.026367
asian-am	0.5321	1.70	0.337512	asian-am	-0.2385	0.79	0.705914
hispanic	-1.0329	0.36	0.030846	hispanic	-0.4037	0.67	0.489713
other	-0.4427	0.64	0.477327	other	-1.3697	0.25	0.022853

5 Discussion of results

5.1 Frequency of tho in the corpus

Results reveal that *tho* has more success as an adverb than as a conjunction. This suggests that *tho* is not semantically equivalent to *though*, as the user-provided definitions of the *Urban Dictionary* and the *Know Your Meme* site suggest. They also show that *though* is 15 times more frequent than *tho*.

Most of the Redditors in the corpus have not yet added *tho* to their linguistic repertoire, with only 296 persons, or 28.4%, using it at least once, which confirms its status as a nonstandard variant. In a small number of cases, however, *tho* may be replacing *though*, or has already replaced *though*, which seems to have disappeared from the CMC vocabulary of 10 Redditors. It is possible that, for these Redditors, *though* and *tho* are semantically equivalent and thus interchangeable.

5.2 Age

The descriptive analysis seemed to indicate an effect of age in the use of *tho*, but it was not confirmed by the regression models, to the exception of the model containing only cisgender females, and only for the oldest group of females. It is surprising that age seems to play such a small role in the use of *tho*, since it has been shown to be correlated with the use of CMC forms such as nonstandard spellings or emoticons (Sánchez-Moya & Cruz-Moya 2015; Oleszkiewicz et al. 2017). This finding shows that inferential methods are necessary to confirm trends revealed by descriptive statistics.

5.3 Race

Findings about race is where effect size was the strongest. The statistical analysis revealed that blacks appear to have adopted *tho* more readily than other Redditors. It may not be completely surprising, given the context *tho* was first used and popularized in. Even if it is difficult to say to what extent it is connected to African American Vernacular English, *tho* gained prominence through a video created by a black man, KingBach, and an *Urban Dictionary* definition, submitted to the site in 2010, links it to Drake and Kid Cudi, two black (mixed race) men (Urban Dictionary: tho, n.d.). If it is indeed a racially marked form, then *tho* and other nonstandard spellings could be some of the linguistic strategies non-whites use to differentiate themselves from the overwhelmingly white Reddit user base.

Tho may also be one of the CMC equivalents of *man*, *bro*, and other African American Vernacular English lexical features which have been appropriated by white and Asian youth (Bucholtz 1999; Chun 2001). It was used by 27.6% of white Redditors, which may show that the nonstandard spelling is in the process of being adopted by white Redditors, maybe for its "coolness factor", and as a way to index masculinity and heterosexuality. Asian American women seem to be those who distance themselves the most from this “cool” spelling of *though*, according to the gender-specific model (Table 10).

5.4 Gender

Effect of gender on the use of *tho* was significant in the model with interactions, which does not take transgender and non-binary people into account. According to this model, black males use *tho* significantly more than white females. The gender-specific models, with only cisgender males and females, showed both similarities and differences between genders. Black males and females were the most likely to use *tho*. Hispanic males also were more likely to use *tho* than white males, but Hispanic females did not behave significantly differently than white women. Asian American females used *tho* less than other groups, but it was not the case of Asian American males. These results show that race and gender should not be considered in isolation, and that interactions should be integrated to regression models, especially in studies that adopt an intersectional perspective. They are able to highlight phenomena otherwise not significant or visible.

6 Limitations

The greatest limitation of this study is perhaps its lack of generalizability. The convenience sample is not representative of Reddit. Also, the presence of outliers in the data may have affected the results of the regression analysis. The models used in this study could benefit from some fine-tuning, to give a more accurate representation of the forces at play and to account for the presence of outliers. The fact that the study takes three variables into account, with five levels in each variable, is also an issue for the statistical analysis, since the number of items per category is sometimes fairly low. The make-up of race categories can also be problematic, especially for the Hispanic category, which may contain users of different racial backgrounds.

7 Conclusion and future work

By focusing on a single linguistic feature, this work has attempted to map out some of the dynamics of race, gender, and age, and the way they manifest in the linguistic choices made online. Further examination of the Reddit corpus and of other corpora is needed to know what other CMC elements black and Hispanic Reddit users have adopted more readily than other ethnicities. It would also help to understand how these nonstandard features make their way across the internet population. A diachronic study of *tho* on Reddit or other platforms may allow to pinpoint who adopted the form first. It also remains to be seen to what extent *tho* remains productive in the years to come, and if more internet users are going to abandon *though* in favor of *tho*.

8 References

Bailey, Cameron. (1996). Virtual Skin: Articulating race in cyberspace. In Mary Anne Moser and Douglas MacLeod (eds.), *Immersed in technology: Art in virtual environments*, 18-24. MA: MIT Press.

Bamman, David, Eisenstein, Jacob, & Schnoebelen, Tyler. (2014). Gender identity and lexical variation in social media. *Journal of Sociolinguistics, 18*(2), 135–160.
Baron, Naomi S. (2004). See you online: Gender issues in college student use of instant messaging. *Journal of Language and Social Psychology, 23*(4), 394-423.
Barthel, Michael, Stocking, Galen, Holcomb, Jesse, & Mitchell, Amy. (2016, February 25). Reddit news users more likely to be male, young and digital in their news preferences. Retrieved October 28, 2017, from http://www.journalism.org/2016/02/25/reddit-news-users-more-likely-to-be-male-young-and-digital-in-their-news-preferences/
Baugh, John. (1983). *Black street speech: Its history, structure, and survival*. University of Texas Press.
Bucholtz, Mary. (1999). You da man: Narrating the racial other in the production of white masculinity. *Journal of Sociolinguistics, 3*(4), 443–460.
Bucholtz, Mary. (2010). *White kids: Language, race, and styles of youth identity*. Cambridge University Press.
Coats, Steven. (2017). Gender and lexical type frequencies in Finland Twitter English. *Studies in Variation, Contacts and Change in English*, 19.
Colley, Ann & Todd, Zazie. (2002). Gender-linked differences in the style and content of e-mails to friends. *Journal of Language and Social Psychology, 21*(4), 380-392.
Cougnon, Louise-Amélie & François, Thomas. (2010). Quelques contributions des statistiques à l'analyse sociolinguistique d'un corpus de SMS. JADT 2010: 10th International Conference on Statistical Analysis of Textual Data.
Crystal, David. (2011). *Internet linguistics: A student guide*. Milton Park, Abingdon; New York, NY: Routledge.
Cutler, Cecilia. A. (1999). Yorkville crossing: White teens, hip hop and African American English. *Journal of sociolinguistics, 3*(4), 428-442.
Dat Tho | Know Your Meme. (n.d.). Retrieved April 28, 2018, from http://knowyourmeme.com/memes/dat-tho
Del-Teso-Craviotto, Marisol. (2008). Gender and sexual identity authentication in language use: The case of chat rooms. *Discourse Studies, 10*(2), 251–270.
Eckert, Penelope, & McConnell-Ginet, Sally. (1995). Constructing meaning, constructing selves. In Kira Hall & Mary Bucholtz (eds), *Gender articulated: Language and the socially constructed self*, 469-507. London and New York: Routledge.
Finlay, S. Craig. (2014). Age and gender in Reddit commenting and success. *Journal of Information Science Theory and Practice, 2*(3), 18–28.
Fought, Carmen. (2002). *Chicano English in context*. Springer.
Gonzalez-Barrera, Ana, & Lopez, Mark Hugo. (2015, June 15). Is being Hispanic a matter of race, race or both? Retrieved May 2, 2019, from https://www.pewresearch.org/fact-tank/2015/06/15/is-being-hispanic-a-matter-of-race-ethnicity-or-both/
Green, Lisa. J. (2002). *African American English: a linguistic introduction*. Cambridge University Press.

Heiden, Serge, Magué, Jean-Philippe, & Pincemin, Bénédicte. (2010a). TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement. In Sergio Bolasco, Isabella Chiari, Luca Giuliano (Ed.), Proc. of 10th International Conference on the Statistical Analysis of Textual Data - JADT 2010) (Vol. 2, p. 1021-1032). Edizioni Universitarie di Lettere Economia Diritto, Roma, Italy. Online.
Herring, Susan C. (1994). Gender differences in computer-mediated communication: Bringing familiar baggage to the new frontier. In V. Vitanza (ed.) *CyberReade*r, 144-154. Allen and Bacon.
Herring, Susan C. (2003). Gender and power in on-line communication. In J. Holmes & M. Meyerhoff (Eeds.), *The handbook of language and gender*, 202-228. Malden, MA: Blackwell.
Herring, Susan C., & Paolillo, John C. (2006). Gender and genre variation in weblogs. *Journal of Sociolinguistics, 10*(4), 439–459.
Herring, Susan C., & Zelenkauskaite, Asta. (2008). Gendered typography: abbreviation and insertion in Italian ITV SMS. In IUWPL7: Gender in language: Classic questions, new contexts, edited by Jason F. Siegel, Traci C. Nagle, Amandine Lorente-Lapole, and Julie Auger (pp. 73–92). Bloomington, IN: IULC Publications.
Hilbe, Joseph M. (2014). *Modeling Count Data*. Cambridge Books.
Huffaker, David A., & Calvert, Sandra. L. (2005). Gender, identity and language use in teenage blogs. *Journal of Computer-Mediated Communication, 10*(2).
Kemp, Nenagh. (2010). Texting versus txtng: reading and writing text messages, and links with other linguistic skills. *Writing Systems Research, 2*(1), 53–71.
Kendall, Lori. (2011). "White and nerdy": Computers, race, and the nerd stereotype. *The Journal of Popular Culture, 44*(3), 505–524.
King, Ritchie, & Olson, Randy. (2015, November 18). How The Internet* Talks. Retrieved April 28, 2018, from https://projects.fivethirtyeight.com/reddit-ngram/
Kolko, Beth E., Nakamura, Lisa, & Rodman, Gilbert E. (Eds.). (2000). *Race in cyberspace.* New York: Routledge.
Jackman, Simon. (2017). pscl: Classes and methods for R developed in the political science computational laboratory. United States Studies Centre, University of Sydney. Sydney, New South Wales, Australia. R package version 1.5.2. URL https://github.com/atahk/pscl/
Labov, William. (1972). *Language in the inner city: Studies in the Black English vernacular* (Vol. 3). University of Pennsylvania Press.
Labov, William. 2001. *Principles of linguistic change. Volume II: Social factors*. Oxford: Blackwell.
Leap, William. L. (1993). *American Indian English*. University of Utah Press.
Lee, Christine. (2003). How does Instant Messaging affect interaction between the genders? Stanford, CA : The Mercury Project for Instant Messaging Studies at Stanford University.
Luor, Tainyi, Wu, Ling-ling, Lu, Hsi-Peng, & Tao, Yu-Hui. (2010). The effect of emoticons in simplex and complex task-oriented communication: An empirical study of instant messaging. *Computers in Human Behavior, 26*(5), 889–895.

Marshall, David. F. (2011). The reforming of English spelling. In J. Fishman & O. Garcia (eds.), *Handbook of Language and Ethnic Identity* (vol 2), 113-125. New York, NY: Oxford University Press.
Massanari, Adrienne. (2017). #Gamergate and The Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society, 19*(3), 329–346.
McCulloch, Gretchen. (2015, May 27). The evolution of "that [noun] though." Retrieved April 19, 2018, from http://mentalfloss.com/article/64323/evolution-noun-though
Ogletree, Shirley M., Fancher, Joshua, & Gill, Simran. (2014). Gender and texting: Masculinity, femininity, and gender role ideology. *Computers in Human Behavior*, 37, 49–55.
Oleszkiewicz, Anna, Karwowski, Maciej, Pisanski, Katarzyna, Sorokowski, Piotr, Sobrado, Boaz, & Sorokowska, Agnieszka. (2017). Who uses emoticons? Data from 86702 Facebook users. *Personality and Individual Differences, 119*(Supplement C), 289–295.
Peersman, Claudia, Daelemans, Walter, Vandekerckhove, Reinhild, Vandekerckhove, Bram, & Van Vaerenbergh, Leona. (2016). The Effects of Age, Gender and Region on Non-standard Linguistic Variation in Online Social Networks.
Prada, Marilia, Rodrigues, David L., Garrido, Margarida V., Lopes, Diniz., Cavalheiro, Bernardo, & Gaspar, Rui. (2018). Motives, frequency and attitudes toward emoji and emoticon use. *Telematics and Informatics, 35*(7), 1925–1934.
R Core Team. (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.
Ranow, George R. (1954). Simplified spelling in government publications. *American Speech, 29*(1), 36–44.
Rao, Delip, Yarowsky, David, Shreevats, Abhishek, & Gupta, Manaswi (2010). Classifying latent user attributes in Twitter. In 2nd International Workshop on Search and Mining UserGenerated Content. ACM.
Romaine, Suzanne. (2003). Variation in language and gender. In Janet Holmes and Miriam Meyerhoff (eds.),
The handbook of language and gender, 98-118. Oxford: Blackwell.
Sánchez-Moya, Alfonso, & Cruz-Moya, Olga. (2015). Whatsapp, textese, and moral panics: Discourse features and habits across two generations. *Procedia - Social and Behavioral Sciences, 173*, 300–306.
Schilling-Estes, Natalie. (2002). Investigating stylistic variation. In J. Chambers, P. Trudgill and N. Schilling-Estes (eds), *The handbook of language variation and change*, 375-401. Oxford: Blackwell.
Sebba, Mark. (2012). In Alexandra Jaffe, Jannis Androutsopoulos, Mark Sebba & Sally Johnson, Sally (eds.), *Orthography as social action: Scripts, spelling, identity and power* (Vol. 3), 1-20. Walter de Gruyter.

Settanni, Michele, & Marengo, Davide. (2015). Sharing feelings online: studying emotional well-being via automated text analysis of Facebook posts. *Frontiers in Psychology*, 6.
Smitherman, Geneva. (1977). *Talkin and testifyin: The language of Black America* (Vol. 51). Boston : Houghton Mifflin.
Squires, Lauren. (2012). Whos punctuating what? Sociolinguistic variation in instant messaging. In Alexandra Jaffe, Jannis Androutsopoulos, Mark Sebba, & Sally Johnson (eds.), *Orthography as Social Action: Scripts, Spelling, Identity and Power*, 289–32). Mouton de Gruyter.
Tagliamonte, Sali A., & In collaboration with Dylan Uscher, Lawrence Kwok, and students from HUM199Y 2009 and 2010. (2016). So sick or so cool? The language of youth on the internet. *Language in Society, 45*(01), 1–32.
tho' | Definition of tho' in English by Oxford Dictionaries. (n.d.). Retrieved April 28, 2018, from https://en.oxforddictionaries.com/definition/tho'
Tho | Definition of Tho by Merriam-Webster. (n.d.). Retrieved April 28, 2018, from https://www.merriam-webster.com/dictionary/tho
Thompson, Dominic & Filik, Ruth. (2016). Sarcasm in written communication: Emoticons are efficient markers of intention. *Journal of Computer-Mediated Communication*, 21, 105-120.
Thomson, Rob, & Murachver, Tamar (2001). Predicting gender from electronic discourse. *British Journal of Social Psychology, 40*(2), 193-208.
Tossell, Chad C., Kortum, Philip, Shepard, Clayton, Barg-Walkow, Laura H, Rahmati, Ahmad, & Zhong, Lin. (2012). A longitudinal study of emoticon use in text messaging from smartphones. *Computers in Human Behavior*, 28, 659-663.
Trudgill, Peter. (1974). *The social differentiation of English in Norwich* (Vol. 13). CUP Archive
Urban Dictionary: tho. (n.d.). Retrieved January 31, 2019, from https://www.urbandictionary.com/define.php?term=tho
Varma, Roli. (2007). Women in computing: The role of geek culture. *Science as Culture, 16*(4), 359–376.
Waseleski, Carol. (2006). Gender and the use of exclamation points in computer-mediated communication: An analysis of exclamations posted to two electronic discussion L
lists. *Journal of Computer-Mediated Communication, 11*(4), 1012–1024.
Witmer, Diane F. & Katzman, Sandra Lee. (1997). On-line smiles: Does gender make a difference in the use of graphic Accents? *Journal of Computer-Mediated Communication, 2*(4).
Y tho | Know Your Meme. (n.d.). Retrieved April 28, 2018, from https://knowyourmeme.com/memes/y-tho
Zeileis, Achim, Kleiber , Christian, and Jackman, Simon. (2008). Regression models for count data in R. *Journal of Statistical Software 27*(8). URL http://www.jstatsoft.org/v27/i08/.

Zentella, Ana Celia. (1997). Growing up bilingual: Puerto Rican children in New York.