ARTIFICIAL OPERATORS: FUNCTION AND USE IN ENGLISH

by

EMILY AE WILLIAMS

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy at
The University of Texas at Arlington
August, 2022

Arlington, Texas

Supervising Committee:
Laurel Stvan, Supervising Professor
Jeffrey Witzel
Ni An

ABSTRACT

ARTIFICIAL OPERATORS: FUNCTION AND USE IN ENGLISH

Emily AE Williams, Ph.D.

The University of Texas at Arlington, 2022

Supervising Professor: Laurel Stvan

This dissertation provides a descriptive account of the use of ‘artificial operators’ on Reddit. I adopt the term artificial operators to refer to written symbols (e.g., #, ™, ©, ®) with overlapping linguistic and metadiscursive properties that are leveraged for a pragmatic effect (e.g., *I love her even though she’s problematic™)*. I employ a mixed-methods approach, using a combination of corpus, experimental, and machine learning methods. Using a 1.2 billion word diachronic corpus of comments from the popular forum website, Reddit, I demonstrate that these operators are used in a small percentage of Reddit communities. Operator usage is therefore often community-specific and provides a useful indicator of the shared repertoire that exists within online Communities of Practice (Lave and Wenger, 1991; Wenger 1998). Operators often function as stance markers, contributing to all three components of stancetaking (evaluation, positioning, alignment). Operators also interact with adjectives by upscaling, or in some cases, downscaling the adjective meaning. Input from social media users indicate that there are perceivable differences between genuine and figurative uses of the operators but do not show perceivable differences between the meaning contributed by the four distinct operators under controlled conditions. Feature importance scores from machine learning models suggest that author-related features are more important than subreddit-related features in modeling operator use. Ultimately, this dissertation shows that artificial operators are a pragmatic resource that authors use to perform a variety of functions, including stance marking, upscaling, and indicating community membership. This work contributes to broader research around pragmatics in computer-mediated communication (CMC) which has shown that authors use CMC cues and other online textual resources not as a replacement for paralinguistic cues, but as new ways to create nuanced and sophisticated meaning.

ACKNOWLEDGEMENTS

This dissertation would not have been possible without the support and guidance of my professors, mentors and colleagues. My supervisor and committee chair, Laurel Stvan, helped me begin this project in the Fall of 2018 during a seminar on discourse markers. Four years later, she is still offering me her continual guidance and support on this idea. I am incredibly grateful to Laurel for always being open, and for allowing me to take this project in so many exciting different directions. I would also like to thank Jeff Witzel for providing invaluable guidance both in and out of the classroom. His statistics class made me a better linguist, and his guidance on experimental design and ordinal logistic regression helped create some of the most exciting results in this dissertation. My final committee member, Ni An, has taught me an extraordinary amount about natural language processing, data science, and machine learning over the past 4 years. She was already a close mentor to me before she joined this committee, and she has been extremely generous with her time while helping me think through the machine learning portion of this dissertation. Lastly, I would be remiss if I did not acknowledge my former classmate Henry Anderson, who acted as a technical mentor and consultant before and during my work on this dissertation. Henry generously provided me with data, performed data cleaning and aggregation, and discussed ideas with me at length. His assistance made this dissertation better, and I am incredibly grateful for it.

This work has been supported by funding from the Department of Linguistics and TESOL through the Nakamura family, and the Dean's Travel Award from the UTA College of Liberal Arts. An earlier version of this work was presented at the Linguistic Society of America Annual Meeting, which would not have been possible without this support.

DEDICATION

This dissertation is dedicated to everyone who supported me, not only during the gargantuan task of completing a dissertation, but during the global pandemic that has defined the last two years. To my first friend in Texas: Shannon – I would not have survived my first year at UTA without you. I will never forget the rollercoaster we were on together that year, and you hold such a dear place in my heart. To Henry – thank you for your relentless desire to help and to share. When I began the wild journey of being a data scientist and a grad student (all while knowing zero Python) you were always willing and eager to share your knowledge. To (the other/cool/French) Emily—meeting you was a turning point for me here in Texas and at UTA. Life has not been easy the past few years, but it has been an unspeakable relief to have someone who always understands exactly what I'm going through. To my current and former squad at AT&T (Prateek, Sherman, Hector, Kx, Ni, Ravi, Jami): getting the chance to be friends with you has been transformative. You taught me how to be comfortable in my own skin. To Dan Amy, my first and favorite office mate, you were a lifeboat for me when I came to UTA. I am so grateful for the mentorship you offered me during that difficult first year. To Dad, Diana, Hunter, and Dylan – you were a huge part of why I chose this PhD program. Coming here to study was a huge leap of faith into the unknown. Now, it is hard for me to even imagine what my life would look like if I hadn't come here: no AT&T, no Bonnie, no shared love of Survivor?! Unfathomable. Lastly to my mom and Jacob: y'all are the best friends I will ever have. I always knew you believed in me, but I also knew that if I quit this program and decided to join the circus, you'd believe in me then, too. Thanks for knowing me better than I know myself.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

Figure Page

LIST OF TABLES

Table Page

CHAPTER 1

INTRODUCTION

The unprecedented volume of language use online has led to a dramatic change in the way that linguistic innovations are delivered, adopted, and spread in computer-mediated communication (CMC). While CMC has been a well-traversed area of scholarly inquiry dating back to the early 1990s, social media has transformed the way language is used online. Sometimes referred to as part of *Web 2.0* (O'Reilly, 2006), social media has led to increased interaction and multimodality in online spaces. Social media texts are being created at a volume and speed which outpaces not only traditional texts but also other genres of CMC media (Deumert, 2015; Benamara et al., 2018). As a result, social media has been credited with facilitating the rapid spread and adoption of linguistic innovations, as well as countless neologisms each year (Herring et al., 2013; Benamara et al., 2018).

With new websites constantly emerging and enabling new kinds of interaction, the boundaries of what is and is not social media are not always clear. Deumert (2015, p. 561), however, defines social media as follows:

> [An] interactive digital media platforms that allow for the creation and sharing of texts, images and other visual content between people. These include Facebook, LinkedIn, Google+, YouTube, Wikipedia and Twitter, but also older text-based applications such as bulletin board services (BBS), discussion groups and the virtual worlds of multi-user dungeons (MUDs), as well as dyadic applications such as texting, email and chatting.

While this definition enables discussion of these sites together, it is important to note that different social media platforms have different characteristics which undoubtedly impact the language employed by their users. Different social media platforms prioritize different aspects of text and have different restrictions in terms of how much users share, the format in which they share, and where they post. Twitter for example has restrictions on length (280 characters), Reddit allows for unlimited text length but requires users to post in community-based subreddits, while Facebook allows for user-specific privacy controls, often requiring users to be ‘friends’ in order to fully access each other’s content. As such, social media data presents a diverse, complex data source with unique features and challenges for research in linguistics.

An important component of working with social media data is assessing where it fits into existing linguistic theory and where it requires modification and development of new theories. One important feature of social media is its ability to foster new types of communities. A number of studies have suggested that digital communities may even constitute Communities of Practice due to the common development of community-specific norms in online spaces (e.g., Zhang & Watts, 2008; Angouri, 2015; Khazraie & Talebzadeh, 2020; Leuckert & Leuckert, 2020). In this dissertation, I propose that an important norm within Communities of Practice is the use of particular CMC cues.

CMC cues have attracted a great deal of scholarly attention. While they were initially analyzed as substitutions for paralinguistic cues in spoken language (Daft & Lengal, 1984; Kiesler et al., 1984; Brennan, 1998), it has since been argued that CMC cues are linguistic innovations (e.g., Herring, 1999; Baron, 2004; Baron, 2009) which enable sophisticated and nuanced communication online.

This dissertation aims to provide an account of an understudied class of such CMC cues: artificial operators. I use the term artificial operators to refer to symbols (#, ™, ©, ®) that have their origins in extralinguistic purposes, but which have taken on linguistic and communicative meaning. Using a mixed-methods approach, I provide a descriptive account of the usage of these operators in CMC as well as evidence of their pragmatic functions. I demonstrate that these operators have shared features, which motivate their study together. I also show that these operators, alongside other community-specific CMC cues, provide uniquely effective examples of the shared repertoire between members of Communities of Practice in online spaces.

1.1 Artificial operators

I adopt the term artificial operators to refer to symbols whose origins are ‘artificial’ in nature, meaning the original purpose was in the organization or clarification of written language. In contrast with orthographic symbols like punctuation which are used to directly indicate grammatically relevant information, the artificial operators I investigate here have origins that are largely nonlinguistic. The copyright symbol, for example, often indicates that the preceding content has been formally copyrighted. Each of these operators has a morphemic quality in that it affixes onto words or phrases which it modifies. While there are likely more artificial operators outside this set, for the purposes of this dissertation I am interested in four artificial operators that occur in social media contexts (#, ™, ©, ®), as in examples 1-5 (operators have been underlined). Throughout this dissertation, the subreddit that examples come from will be italicized underneath the example, or in the case of blocks of examples from the same subreddit, stated in the prose.

1. recite your oath to **freedom©**
   r/newzealand

2. **The Gay Agenda™** is thriving
   r/rupaulsdragrace

3. Divinity should be fixed **soon®** so just hope for that
   r/DestinytheGame

4. Wow, this is so #sad.
r/im14andthisisdeep

5. OF COURSE it is a **florida man®©™**
r/iamatotalpieceofshit

In crafting a preliminary definition for artificial operators, I make the following assertions about the operators in examples 1-5. Each of these operators is ‘artificial’ in nature; they are symbols which have origins in nonlinguistic purposes. In this dissertation, I refer to artificial operators which appear to be used for these original purposes as genuine, and their non-genuine uses (as in examples 1-5) as figurative. The operators affix onto words (or entire phrases) and cannot stand alone in the sentence. I refer to these words or phrases as affixed units. The operators do not change the syntactic structure of the sentence; removal of the operator yields a sentence with identical syntactic structure. The operators similarly do not alter the syntactic category of the word or phrase onto which they attach. The operators may occur independently (examples 1-4) or together (example 5). Finally, I assert that these operators contribute meaning to the sentence.

As previously stated, each of these operators contributes nonlinguistic meaning. For hashtags on social media, the affixed unit typically represents a clickable hyperlink that takes the user to a page where posts with the same hashtag are aggregated in real time. Hashtags are formed by combining the hashmark (#) with a continuous text string (e.g., #sad). Hashtags may not contain spaces or punctuation; in order for a hashtag to scope over multiple words, they are joined together with word boundaries often denoted by capitalization (e.g., #MeToo). Clicking on a hashtag on social media sites such as Twitter, Instagram, or Facebook will navigate to a page of all posts containing that hashtag. Hashtags therefore are somewhat organizational: they allow social media posts on a single topic to be collected into a single place, enabling large-scale, real-time interaction among a potentially infinite number of users. Finally, the hashtag is notably distinct from the rest of the AOs in this study in that it precedes the text to which it affixes.

The other artificial operators I discuss here, the trademark symbol, copyright symbol, and registered trademark symbol, have their origins in legal contexts, indicating legal existence or ownership by a particular entity (e.g., *Intel*®). In both genuine and figurative uses, they are capable of scoping over multiple words, which may be continuous (e.g., *NiceGuy*™) or separated by spaces (e.g., *Cultural Appropriation Light*©). Individual words may be capitalized, but are not required to be (e.g., *certified hood classic*®). Because of this, the intended scope of these operators is often ambiguous. Finally, in contrast with the hashtag, these operators follow the text to which they affix.

Williams (2021) argued that the hashtag and the trademark symbol have a number of overlapping metadiscursive functions including evaluative metacommentary, ambient affiliation, and critique. In this dissertation, I build upon this work by expanding the set of operators to include the registered trademark symbol and the copyright symbol. This is motivated by the fact that they also behave in strikingly similar ways, as demonstrated by examples 6-9.

6. The shade at **#whitepeople** lmao
   r/rupaulsdragrace

7. You're telling me **White People™** colonized half of the world, killed a ton of people for spices and their food still ended up bland?
   r/BlackPeopleTwitter

8. Never fear minorities! Well-to-do **White People©** are here to be offended for you!
   r/creepy

9. There's a great show called Community where this feature of **White People®** is indicated
   r/news

As evidenced by the examples above, not only do these four operators share the previously described features of artificial operators, but they are capable of being affixed to identical words or phrases. These shared contexts bolster the case for studying these operators together and exploring the degree to which they occur in shared contexts and the degree to which their functions overlap. The use of these symbols beyond their original function raises questions about their purpose and the linguistic functions they serve.

1.2 Research questions

Artificial operators (AOs) are understudied, and therefore present an opportunity for building up an (as of yet absent) descriptive account of their use. Pragmatic and sociolinguistic theory provide an avenue for understanding the contextual meaning contributed by AOs. Central to questions of contextual meaning are whether it is necessary or accurate to distinguish between genuine and figurative uses, and whether or not there are substantial differences between any of the operators. Additionally, the digital nature of social media provides an opportunity to study AOs at scale by collecting a massive corpus of online discourse. Recent advances in computational methods, machine learning, and natural language processing also provide a rich set of tools and approaches for analyzing such massive corpora. With these motivators in mind, the four research questions I answer in this dissertation are below.

> RQ1: What are the linguistic features that condition the use of artificial operators? How have these changed over time? What motivates their use?
>
> RQ2: What are the pragmatic functions of artificial operators as used on Reddit?
>
> RQ3: What are the differences in meaning contributed by ‘genuine’ vs. ‘figurative’ artificial operators? What are the differences in meaning contributed by different artificial operators?
>
> RQ4: Can we predict AO use on Reddit with machine learning models? Will author-related features or subreddit-related features be more important to the models?

As social media data is widely available in large quantities, the broad framework of Corpus Linguistics is an appropriate methodology for RQ1. In recent years, corpora have been used to support studies in fields such as theoretical semantics (e.g., Chen, 2020; Kotowski, 2021), pragmatics (e.g., Stvan, 2006; Browning, 2017), sociolinguistics (e.g., Leuckert & Leuckert, 2020), and discourse analysis (e.g., Taylor, 2014; Fleckenstein, 2019). In performing the corpus analysis, I propose that analyzing these operators through the lens of Communities of Practice provides insights into what motivates their use. I ultimately argue that, as has been found in prior studies, many subreddits constitute Communities of Practice. I furthermore argue that AOs, and other CMC cues like them, provide excellent examples of the shared repertoire that occurs within Communities of Practice.

To approach RQ2, I adopt stance theory (Du Bois, 2007) as a useful framework for discussing meaning within Communities of Practice. Shared ideologies and alignment regarding evaluation of stance objects is an important element of defining community membership. I furthermore highlight the importance of shared community knowledge by focusing on discourse-oriented adjectives and how they interact with AOs.

The qualitative analysis relies on researcher intuitions and is susceptible to researcher bias. To mitigate this, for RQ3, I collected survey data in order to explore the perspectives and reactions of social media users to these operators. This allows exploration of the differences between genuine and figurative operators, as well as the differences among the operators. Lastly, for RQ4, I perform machine learning modeling to see which features are the most important in predicting author usage of AOs.

1.3 Outline of the dissertation

In this chapter, I have provided background on social media as a source of linguistic 'big data.' I have proposed the existence of a class of AOs that are being repurposed in social media contexts in strikingly similar ways. The remainder of this dissertation will be organized as follows: in Chapter 2, I will review the literature on pragmatics research in CMC, as well as past explorations of the hashtag, connecting the research questions I seek to answer to previous work in the field. In Chapter 3, I detail the information about the corpus assembly and overview the quantitative results of the corpus analysis. In Chapter 4, I perform close analysis of excerpts from the corpus, proposing AOs as stance markers. In Chapter 5, I discuss the interesting interactions between AOs and discourse-oriented adjectives, proposing contextually driven scalar modification as another function of some AOs. In Chapter 6, I discuss the two surveys I administered and the results of the data analysis. In Chapter 7, I discuss the two machine learning tasks I completed, which a focus on feature importance as a potential proxy for understanding the motivations underlying AO use. In Chapter 8, I summarize the results of these various efforts and propose conclusions about the meaning contributed by AOs and make recommendations for areas of future work.

CHAPTER 2

LITERATURE REVIEW

As social media becomes increasingly accessible, the sheer volume of linguistic data being produced on the internet is unprecedented in scope and size. Despite concerns about how to conceptualize and discuss the usage of text-based language online, so-called ‘internet linguistics’ (Crystal, 2001) is a growing area of research. While language use online has been referred to in a number of different ways (*netspeak* – Crystal, 2001; *Interactive Written Discourse* – Ferrara et al., 1991; *Online Written English* – Collister, 2011), for the purposes of this project I adopt the term Computer-Mediated Communication (CMC), which refers to language use mediated by the internet via phones, computers, or other networked devices.

Early work in CMC was concerned with where to place CMC in relation to conventional linguistic approaches. There have been those who have discussed whether CMC constitutes a new dialect of language (e.g., Crystal, 2001; Collister, 2011) as well as arguments that CMC constitutes a new register (e.g., Ferrara, et al., 1991; Tagliamonte & Denis, 2008). Much of the early work in this area focused on how CMC relates to the traditional binary of written and spoken language (e.g., Ferrara et al., 1991; Maynor, 1994; Werry, 1996; Yates, 1996). While the degree of ‘spokenness’ of CMC remains a topic of open discussion, there has largely been a consensus that CMC contains a blend of features of both spoken and written language, which vary across the numerous modes of communication within CMC. McSweeney (2018, p. 24) summarizes that CMC is “sufficiently different from spoken and written modalities to treat it as its own language form with systematic and quantifiable norms and conventions.”

The aforementioned differences that make CMC distinct are derived from a number of unique features of CMC—which in contrast with spoken language may be synchronous, asynchronous, or semi-synchronous—as well as cues often associated with CMC such as nontraditional spelling and punctuation, emoticons and emojis, and the widespread use of initialisms and abbreviations. Some have argued (Daft & Lengal, 1984; Kiesler et al., 1984; Brennan, 1998) that these unique features are a result of CMC being an 'impoverished' medium and that these CMC cues are an attempt to substitute for paralinguistic cues that occur in spoken language. Other researchers (e.g., Herring, 1999; Baron, 2004; Baron, 2009) have argued that CMC cues are not merely a replacement for paralinguistic cues, but instead are linguistic innovations which are enabled by features of the CMC medium that foster creativity and language play. The norms and conventions within CMC are not taught. Rather, they are largely established by the users. While there are certainly features regularly associated with CMC, they are not used universally; different CMC norms and practices form within distinct communities (Baron, 2004; Deumert & Masinyana, 2008). CMC has been shown to participate in the same sociolinguistic variation as spoken language, with much of the research in CMC able to "situate itself within established approaches in socially oriented linguistics" (Androutsopoulos, 2011, p. 277). Use of CMC cues, for example, have been shown to vary based on gender (Witmer & Katzman, 1997; Huffaker & Calvert, 2005; Fox et al., 2007) and age (Siebenhaar, 2006; Argamon et al., 2007; Goswami et al., 2009). This has led for some to propose that the sociolinguistic framework of Communities of Practice (CoPs) may be extensible to CMC contexts.

In the remainder of this chapter, I provide an overview of the existing scholarship that is relevant to the goals of this dissertation. In section 2.1, I overview the sociolinguistic framework of Communities of Practice (CoP) and discuss past studies involving CoPs in CMC contexts. In section 2.2, I talk about the body of work dealing with CMC cues, highlighting the contributions these cues make to existing linguistic theory. In section 2.3, I discuss studies which apply existing pragmatic theories to CMC datasets and relate this specifically to my adoption of stance theory and discourse-oriented adjectives as frameworks which effectively capture AO functionality.

2.1 Communities of Practice in CMC

Communities of Practice (CoP) are a sociolinguistic framework (Lave and Wenger, 1991; Wenger 1998) which "characterize membership as being created and maintained through social practices (linguistic or otherwise) at a local level, rather than global categories being imposed on individuals" (Davies, 2005, p.557). Three requirements have been established to deem something a community of practice: mutual engagement, a jointly negotiated enterprise and shared repertoire (Meyerhoff & Strycharz, 2013).

In recent years, a number of studies have emerged which study online CoP (e.g., Zhang & Watts, 2008; Angouri, 2015; Khazraie & Talebzadeh, 2020). Leuckert & Leuckert (2020) specifically propose that subreddits on Reddit may act as CoP by looking at the behavior of users from 3 subreddits and demonstrating how those subreddits relate to the criteria of joint enterprise, mutual engagement and share repertoire. However, one thing I would like to point out is that in this study, they argue for shared repertoire by showing that the most frequent words in each corpus are thematically linked to the topic of the subreddit. While this is helpful, I propose that AOs, because they are niche and highly specialized in meaning, present a compelling opportunity to bolster the argument for subreddits as communities of practice, and for studying digital CoP.

2.2 CMC cues

In addition to explaining the impetus behind the usage of CMC features, there has been a great deal of research attempting to outline the linguistic features of distinct CMC mediums (text message, chatting, blogging, social media) and how they relate to questions of variety and register (e.g., Tagliamonte & Denis, 2008). Androutsopoulos (2011, p. 281) notes that CMC is

> characterized by processes of multimodality and multiauthorship: their content is produced by multiple participants, simultaneously and in part independently of each other; and they host and integrate complex combinations of media and semiotic modes

Investigation into CMC cues themselves has revealed that their use is systematic and rule-governed (Uygur-Distexhe, 2014; Crystal, 2014; McSweeney, 2017) and that participants employ these cues strategically, often expending additional effort to ensure that their communication is written as they intended (McSweeney, 2018). Collister (2011) explored the importance CMC cues in repair strategies in CMC, noting that CMC has developed unique repair strategies which do not have direct correlates in spoken language. Vandergriff (2013) similarly found that CMC cues may not be merely translated into some sort of spoken equivalent. Instead, context plays a critical role in deriving the meaning of CMC cues. McSweeney (2018, p. 24) argues that

> the key thing that differentiates digital communication from either written or spoken language is what these features do for the digital writer and how they have come to take on a variety of pragmatic meanings within a conversation.

Understanding the functions of CMC cues has often been at the center of CMC research. The use of nonstandard orthography in CMC has been shown to be intentional stylistic choices or indicators of closeness (Thurlow & Brown, 2003; Negrón Goldbarg, 2009; Tagg, 2009). Zappavigna (2012) argues that nonstandard orthography may be employed for ‘upscaled graduation’ to upscale the interpersonal meaning of a sentence. Heath (2017) similarly argues that non-standard orthography may be used for emphasis as nonstandard orthography will attract more attention from the reader. Emoticons and emojis have also attracted a great deal of scholarship. Markman and Oshima (2007) argue that the function of emoticons is analogous to punctuation. Baron (2009) note that emoticons are not merely replacements for facial expressions but have highly context-dependent meaning. According to Baron (2009, p. 14), "emoticons are no more univocal than are words in ordinary language, and therefore cannot be assumed to unambiguously clarify user intention or emotion." More recent work has shifted from emoticons to the more popular emojis, which similarly have been found to be used as "multimodal affective markers" which rely heavily on context for meaning (Na'aman et al. 2017). The importance of context-dependence as well as the lack of clear speech equivalents have demonstrated arguments against the idea that CMC cues are 'needed' to clarify meaning. Rather, CMC cues such as emoticons and nonstandard orthography are examples of users taking advantage of an abundance of linguistic resources in CMC to clarify and express meaning with as much nuance as in spoken language.

2.3 Past work on artificial operators

In recent years, the hashtag has emerged as another critical element within CMC that is worthy of linguistic analysis. Huang et al. (2010) analyzed the functions of hashtags on Twitter, demonstrating that they are often used to ensure content appears in the correct informational stream. Zappavigna (2012) similarly argued that hashtags are primarily used to organize and catalog tweets, making them findable for other Twitter users. This searchability has led for some to claim that hashtags enable community formation (Bruns & Burgess, 2011; Yang et al., 2012). Page (2012) notes that hashtags can lead to increased user participation while Rossi and Magnani (2012) similarly argue that this searchability facilitates highly focused interactions in which a massive number of users are able to participate in conversations around a single topic. In summary, the hashtag's function to enable search and aggregation of tweets around a similar topic was central to much of the early work around the hashtag. The hashtag has also been shown to play a role in information management (Kehoe & Gee, 2011; Zappavigna, 2015), with hashtags often being used as labels to denote what a particular text is about. Browning (2017) studied the pragmatic functions of hashtags, framing them as discourse markers. While some of these functions are not linguistic, per se, they are related to linguistic (and pragmatic) concerns, particularly how the hashtag may facilitate new types of conversations and may function to emphasize certain topics within a particular text.

Crucially, the research on the hashtag described thus far has taken as its object the entire hashtagged unit (e.g., #MeToo) as opposed to considering the effect that the hashtag operator (#) has on the phrase to which it affixes (*MeToo)*. There has been little work analyzing the pragmatic and linguistic effects of the hashtag when it is used outside contexts such as Twitter, where it carries out the function of making text searchable.

Finally, very little has been said about other artificial operators (AOs) which share similar qualities to the hashtag (e.g.,™, ©, ®). Williams (2021) compares the uses of the hashtag and trademark on Reddit, is the first and only study that proposes studying these operators together. In this study, I argue that the trademark symbol performs many of the same functions as hashtags. I follow Zappavigna in claiming that hashtags contribute metadiscursive meaning and create two orders of meaning: tagged and untagged text. Zappavigna (2018) argues that the hashtag participates in construing evaluation, creating ambient affiliation, and enacting criticism. In Williams (2021), I argue that the trademark also performs these functions. There has yet to be any work performed on the entire group of AOs I take an interest in for this dissertation.

2.4 Pragmatics and CMC

As the scope of CMC research continues to expand, it is of critical importance to tie analyses of the CMC context back to pragmatic theory. This will enhance understandings of CMC itself, as well as enriching pragmatic theory to account for the role of contextual meaning in (all types of) language use. A number of CMC studies have successfully employed 'traditional' pragmatic frameworks to analyze CMC cues. Dresner and Herring (2010), for example, analyzed emoticons using speech act theory (Searle & Searle, 1969), arguing that emoticons are often used as indicators of illocutionary force. Dresner and Herring used their analysis of emoticons as speech acts to argue that emoticons should be viewed as part of the text and as part of linguistic (as opposed to nonlinguistic) behavior. By demonstrating the efficacy of using an existing pragmatic framework to analyze emoticons, Dresner and Herring reveal that CMC cues are not fundamentally different from other more familiar types of language use.

Wikström (2014) marks another study which employed a traditional pragmatic analysis to look at a CMC cue, adopting a Gricean and speech acts framework to analyze hashtags. Wikström employed this framework to allow for analysis that, unlike much of the previously discussed work on hashtags, is truly linguistic in nature, focusing on the hashtag's communicative functions as opposed to its organizational ones. Wikström argues that hashtags are often used as ways for speakers to flout the traditional Gricean maxims. Wikström demonstrates the efficacy of using traditional Gricean framework, as well as a traditional speech acts framework, to analyze CMC.

McSweeney (2018) also enables her analysis of CMC by placing it in relation to Grice's cooperative principle. McSweeney importantly frames violations of spelling convention as 'flouting' the hypothetical maxims of the CMC context, noting that with tools like autocorrect, nontraditional orthography often requires more effort to violate or flout. She advocates for the usefulness of using such a traditional Gricean approach to understand CMC, but also notes that CMC may have its own new maxims and rules.

Each of the above studies demonstrated that CMC may be effectively studied and analyzed using an approach grounded in traditional pragmatic theory. The traditional Gricean framework provided by the cooperative principle provides useful baseline understandings with which to explore the function and intention behind a number of CMC cues. However, as McSweeney notes, the maxims and rules which govern CMC may be different than those which are traditionally discussed with relation to spoken language.

2.5 Conclusion

The linguistic innovations which are rapidly evolving on the internet comprise an area of research that demands a more complete analysis. Furthermore, development of a linguistic theory that fully accounts for the unique features of CMC is an effort that is still very much underway. Therefore, it is the goal of this study to utilize corpus approaches in combination with qualitative and experimental analysis to propose where AOs fit within existing pragmatic theory as well as how they relate to the sociolinguistic concept of CoPs. In this way, this dissertation seeks to fill two gaps. The first is to present an account of AOs in CMC, an area that has attracted little scholarly attention. The second is to leverage the analysis of these AOs to contribute to the developing linguistic theory of CMC.

CHAPTER 3

CORPUS ANALYSIS

In chapter 2, I provided background on previous work done in CMC around artificial operators (AOs). I also proposed the sociolinguistic theory of Communities of Practice (Lave and Wenger, 1991; Wenger 1998) as a useful framework for understanding the motivations behind AO usage. In this chapter, I present the results of the quantitative corpus analysis. While the corpus described here underpins all the work of this dissertation, including the qualitative analyses, in this section I focus on describing the automated and quantitative analyses performed on it. In section 3.1, I provide background on the framework of Corpus Linguistics, which underlies the analysis in chapters 3-7. In section 3.2 I provide background on the corpus collection and data processing. In section 3.3 I present the results of the analysis performed over the entire corpus, with a focus on changes in frequency, affixed unit length, and part of speech information over time. In section 3.4, I present a case study of three subreddits, performing a comparative analysis of three distinct communities where AO use is relatively common. In section 3.5, I summarize the findings and outline key insights.

3.1 Corpus Linguistics

Corpus Linguistics (CL) is a methodological approach which relies on the use of software and programming tools to perform quantitative analysis of bodies of text (Evison, 2010). These bodies of text are said to be corpora if they are:

> …a body of naturally occurring texts that is (a) representative of a specified type of language; (b) relatively large in terms of word count; and (c) machine-readable. (Fitzsimmons-Doolan, 2015, p. 107)

One of the main benefits of using a corpus is the fact that the language is naturally occurring and therefore may be more indicative of ‘real’ language use. With increasingly powerful software tools available, corpus approaches present a relatively fast method for dealing with large quantities of data (Mautner, 2009).

Some have criticized corpus methods, noting that corpora are decontextualized bodies of text that only represent a ‘partial account of real language’ (Widdowson, 2000, p. 5). Furthermore, since the quality of a corpus is central to the research outcomes, a poorly constructed corpus may lead to less reliable outcomes for corpus studies. For example, a corpus may not be sufficiently representative of the features it is used to study (McEnery et al., 2006), or a corpus may be biased toward a particular kind of language use (Baker et al., 2008). It is important to keep these limitations in mind while constructing and analyzing a corpus.

3.2 Reddit corpus

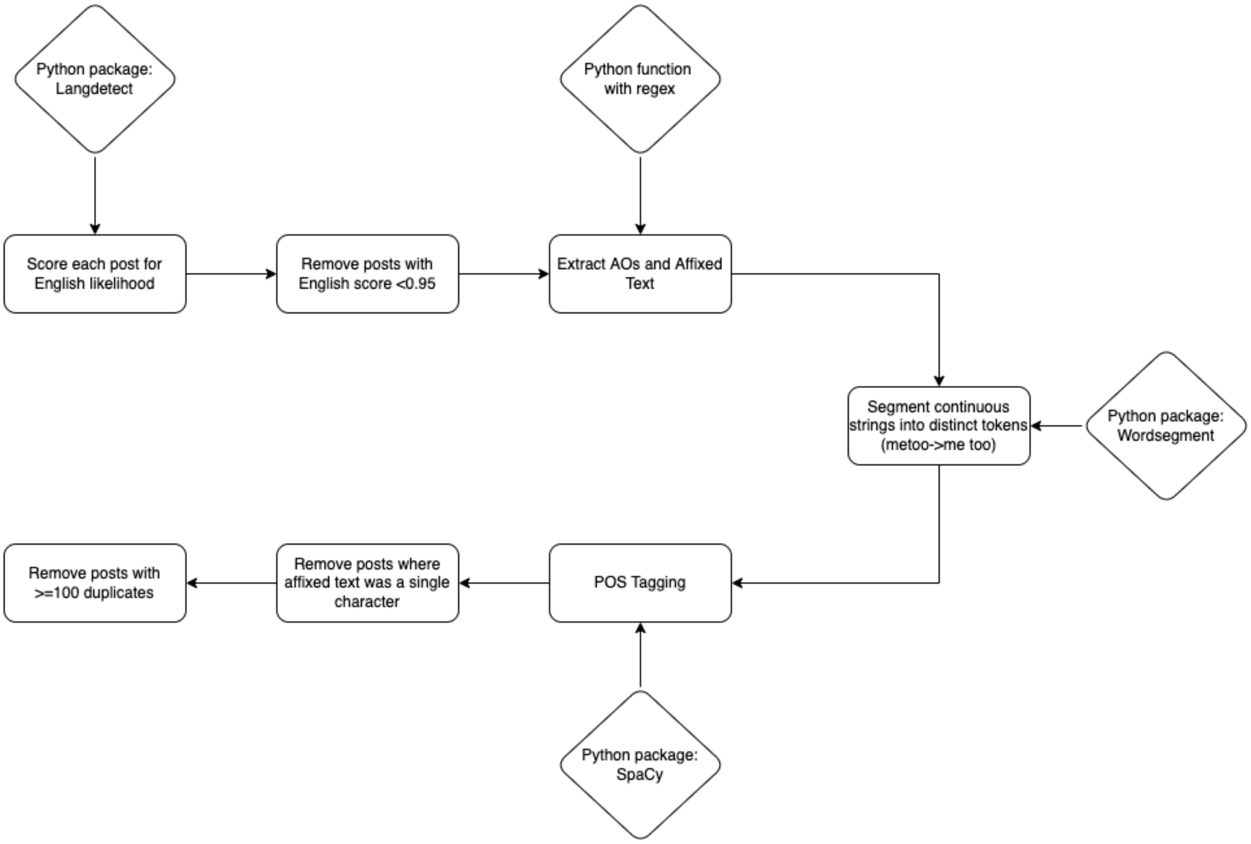
Reddit was determined to be an optimal dataset for this project for a number of reasons. The Pushshift Reddit dataset is a publicly available dataset spanning more than 15 years, allowing observations of changes in language use over time. Reddit is organized into subreddits, which have a central theme or topic. The organization of text into subreddits allows for the separation of data into clearly defined communities to quantify the potential influence of in-group knowledge or community norms. Reddit is distinct from social media sites such as Twitter, Instagram, and Facebook in that the use of a hashtag does not create a clickable hyperlink or enable any special behavior whatsoever; technologically, it behaves no different from any other typed character. That is, hashtags on Reddit do not serve the function of aggregating similar posts. In this sense, Reddit data avoids the possibility that users are employing hashtags for the purpose of starting a trend or increasing the findability of a comment.

On Reddit, users share posts in subreddits, which are organized around a particular topic. Within each subreddit, discussions are further divided into ‘threads,’ consisting of one original post, and a series of responses. Users can respond directly to the original post, or to any response, potentially creating many parallel chains of responses. The Pushshift Reddit dataset (Baumgartner et. al., 2020) is a collection of all such publicly visible comments. Users can request that their data be removed from the dataset, but barring those requests, the dataset is comprehensive, containing even data from now defunct or banned subreddits. The entire corpus was queried via Google Big Query (Fernandes & Bernardino, 2015), which hosts a copy of the corpus available for public use. This copy of the dataset contains all Reddit comments from December 2005 to December 2019 (barring aforementioned requests for removal).

Throughout this chapter I will refer to three datasets. The full corpus refers to the entire collection of Reddit comments from 2005 to 2019, with no filters applied. The AO dataset refers to a collection of only those posts which contain AOs. Lastly, I will refer to a case study dataset which contains all comments from three specific subreddits.

The full corpus was queried to extract posts with AOs for the AO corpus. Posts which contained #, ™, ©, or ® were included in the AO corpus. In the case of the ™, © and ® only the Unicode symbols were used, to avoid false positives that might have been brought in by the letters *TM*. For hashtags, I included a space in front of the hashtag in the query to avoid pulling those posts which might have a hashtag inside of a URL. After these posts were extracted, data underwent preprocessing according to the steps in Figure 3.1 below.

Figure 3.1: Preprocessing flow


Python package: Langdetect
Python function with regex
Score each post for English likelihood
Remove posts with English score <0.95
Extract AOs and Affixed Text
Segment continuous strings into distinct tokens (metoo->me too)
Python package: Wordsegment
Remove posts with >=100 duplicates
Remove posts where affixed text was a single character
POS Tagging
Python package: SpaCy

As detailed in the preprocessing diagram, posts in the AO corpus were scored for their likelihood of English using the Python package Langdetect (Shuyo, 2010). Posts which scored >0.95 were included in the AO corpus. This ensures a higher precision in identifying posts that are definitely English, as opposed to something like Dutch. Such an aggressive cutoff also helps mitigate false positives from very short texts. Next, AOs and the affixed text associated with them were automatically extracted using a regular expression (regex). Regular expressions are ways to specify complex text patterns, which can then be used to search for matching strings or documents. The regex looked for one of four conditions:

1. A hashtag followed by text at the beginning of a comment (e.g., *#First I want to say…*)

2. A hashtag followed by text with a space in front of the hashtag (e.g., *Yeah, she is part of the #maga crowd*)

3. Sequential capitalized text in front of ©/™/® with an optional space in between the text and the operator (e.g., *don't be a Nice Guy* ™)

4. Text immediately preceding ©/™/® (e.g., *I'm lost*®)

As I mentioned briefly in Chapter 1, scope is often ambiguous with AOs, particularly ©/™/®. In analyzing the corpus, it is impossible to say with certainty what the intended scope of the operator was, even upon manual inspection. However, since capitalization is often used to denote scope, for the purposes of automatic extraction this was deemed sufficient. This means that in the case of lower-case affixed text followed by a ©/™/®, the regex would always extract a single word, potentially biasing the dataset toward short affixed units. However, since there is no way to know the author's intended scope, even with manual inspection, this approach was seen as acceptable. After extraction of the AO, the affixed unit was passed to a Python package, Wordsegment (Jenks, 2018), which automatically separates words that co-occur without any spaces between. Next, the segmented text underwent part of speech tagging, which I detail in section 3.3.3. Next, I removed comments in which the AO was a single letter to remove noise from the data. Finally, I removed duplicate comments if they occurred more than 100 times in the corpus. The goal of this was to avoid scenarios where a templated comment made by a bot or spam account would have undue influence on the results of the analysis.

In Williams (2021), I drew strict distinctions between the genuine usage of these symbols (e.g., *2019 HEARTHSTONE® GRANDMASTERS OFFICIAL COMPETITION)* and their communicative or figurative use (e.g., *Blizzard: Yea we are gonna fix it soon*®). While the focus of this work is on understanding the pragmatic and sociolinguistic functions of the figurative operators, it is not always possible to clearly distinguish between the genuine and figurative uses.

I discuss this issue in more detail in Chapter 6. As such, for the corpus analysis, all instances of the symbols were counted, regardless of context. This allowed for an analysis of overall symbol use, while in the following chapters, I spend more time looking specifically at the figurative use.

3.3 Results: AO dataset

In this section, I describe the corpus analysis that I did to chart dynamic changes in the use of AOs over time. After preprocessing, the AO dataset contained 5M posts, 8.4M AOs, and 1.1B words. The analysis contains three main components: frequency analysis, affixed unit length analysis, and part of speech tagging.

3.3.1 Frequency analysis

Frequency analysis provides an excellent entry-point into corpus analysis. As this is a diachronic corpus, one of the goals was to explore how, if at all, frequency in AOs has changed over time. 2005 data is not included in the plots in this section because there is only one month of data from that year. Unless otherwise noted, Figure 3.2 and the other plots in this chapter have y-axes on a logarithmic scale, due to the vast frequency disparity between hashtags and the other operators. To preserve readability, numeric values are not included on many plots in this chapter. However, in cases where the plots do not have numeric labels, the underlying values are available in Appendix A.

Figure 3.2: AO comments over time (raw frequency)

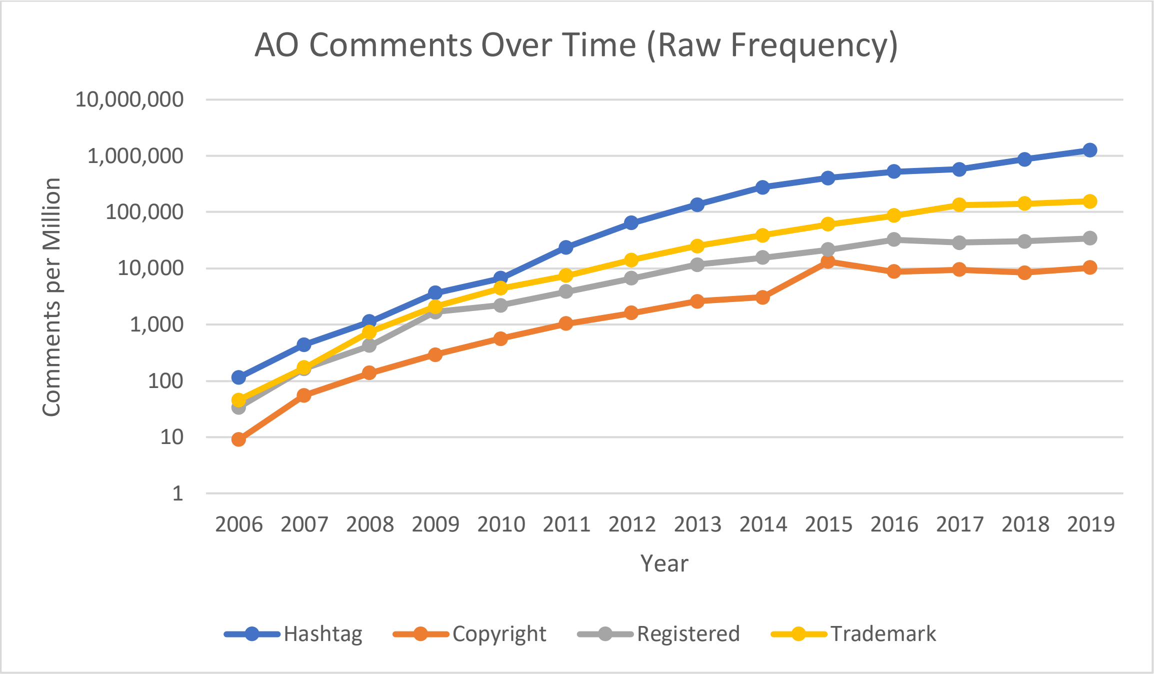
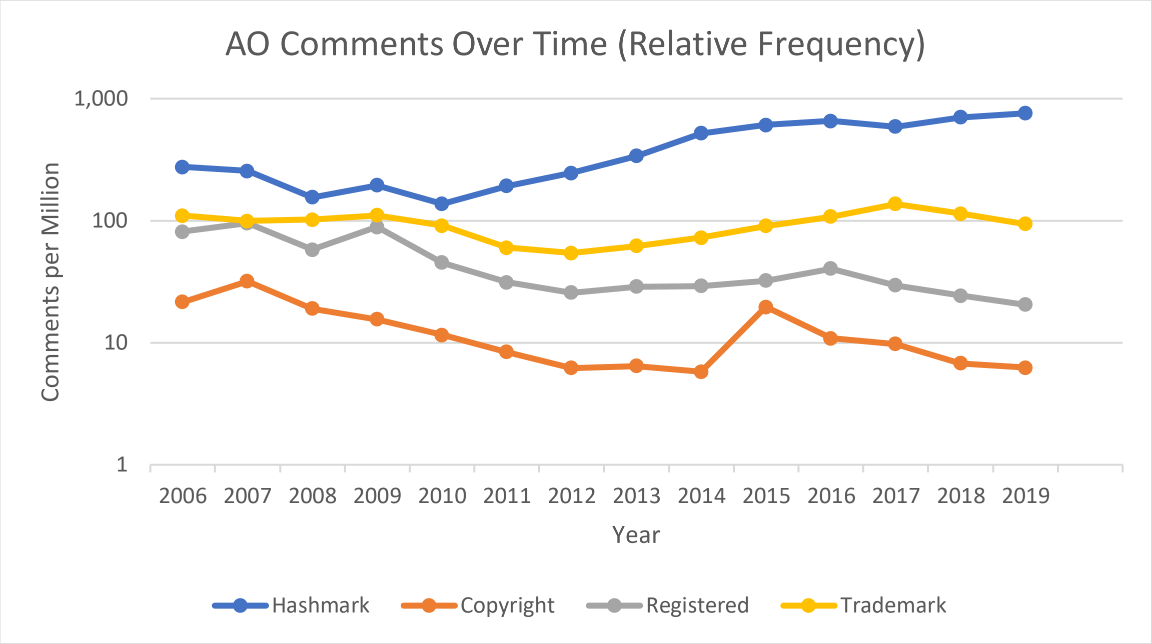


Figure 3.2 above displays the raw frequency of comments containing each symbol from 2006 to 2019. The raw frequency of comments for each of the AOs has increased more than a thousandfold. However, this is unsurprising as Reddit itself has grown exponentially during the same time frame. As such, relative frequency of comments in comments per million (CPM) are presented in Figure 3.3 below.

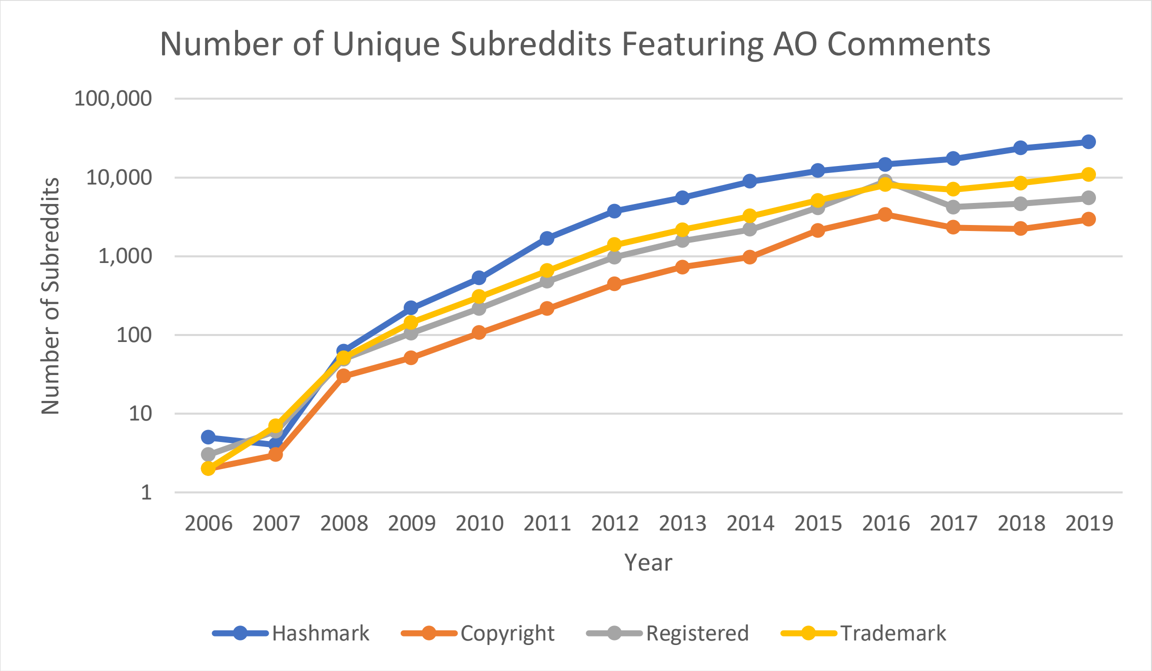
Figure 3.3: AO comments over time (relative frequency)



The relative frequency of the hashtag has shown an overall increase, particularly after the year 2010. In 2006 the hashtag had a relative frequency of 275.7 CPM and by 2019 that had increased to 758.8 CPM. Other operators, however, have not shown similar increases in usage. The trademark symbol has been relatively stable across the time period of the corpus with 110.2 CPM in 2006 and 94.0 CPM in 2019. The copyright symbol and the registered trademark have actually seen declines compared to their relative frequency in the first 5 years of the corpus. The registered trademark symbol dropped from 81.5 CPM in 2006 to 20.5 CPM in 2019 and the copyright symbol dropped from 21.6 CPM in 2006 to 6.2 CPM in 2019. Therefore, it cannot be stated that AOs, in general, are on the rise on Reddit, with the exception of the hashtag, which has trended upward in relative frequency. It is possible that in the cases of the copyright and registered trademark, usage was inflated or anomalous due to the small sample size of posts in the first few years of the corpus.

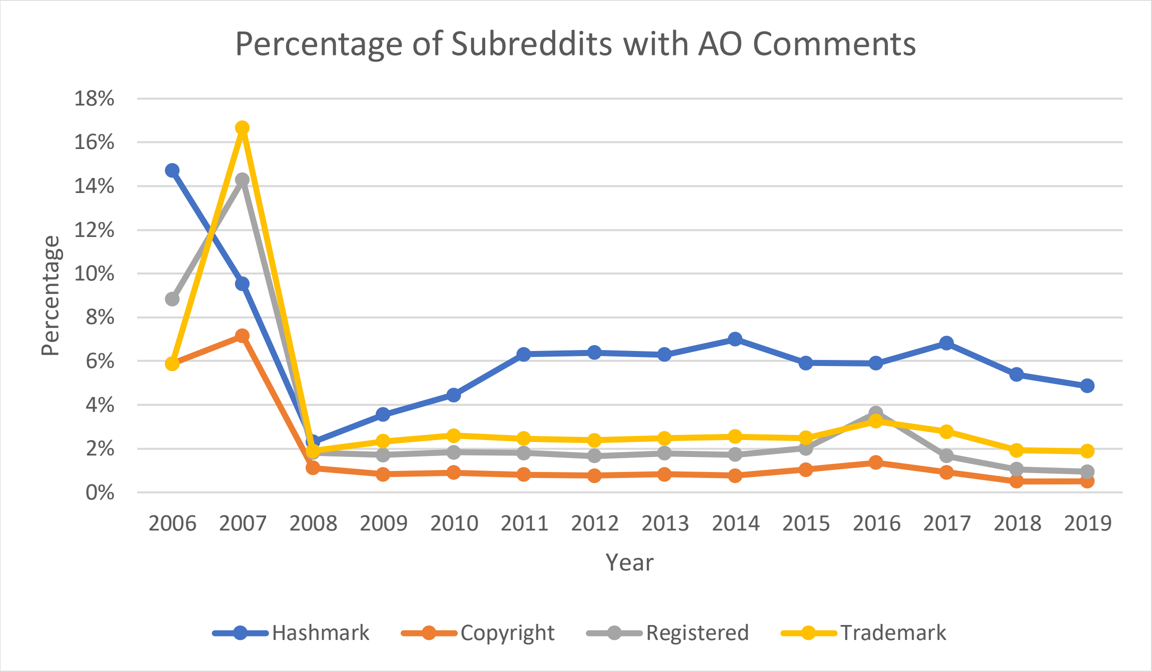
Another way of gauging the use of AOs on Reddit is by looking at the number of distinct subreddits that contain comments with AOs. This is shown below in Figure 3.4.

Figure 3.4: Unique subreddits with AO comments



As with the raw frequency of the AOs, the number of subreddits featuring AOs steadily increased as Reddit grew. The hashtag again has the highest dispersion across subreddits, and the copyright has the lowest. However, in order to contextualize this, Figure 3.5 below (non-logarithmic axis) shows the percentage of subreddits which contain AOs.

Figure 3.5: Percentage of subreddits with AO comments



The first three years of the data—2006 through 2008—should be interpreted with caution, due to the extremely small sample size of subreddits that existed during these years. There were 34 total subreddits in 2006 and 42 in 2007. This number shot up to 2695 in 2008, which co-occurs with the sudden drop off for all the operators. However, the broader trend that we have seen in the frequency analysis persists here: the hashtag has the highest relative frequency and the copyright has the lowest. Regardless of these differences among the operators, the results here suggest that AO usage is not a defining characteristic of most subreddits. Instead, the percentage of subreddits where AOs are used represents a very small percentage of overall communities.

In summary, raw frequency of all 4 AOs has increased substantially over time both in terms of volume of comments and volume of subreddits where those comments occur. However, relative frequency has increased only for the hashtag. These findings suggest that—contrary to the hypothesis that these operators are increasing in use—they continue to represent a highly specialized behavior that is not necessarily representative of overall Reddit language use.

3.3.2 Affixed Text Length

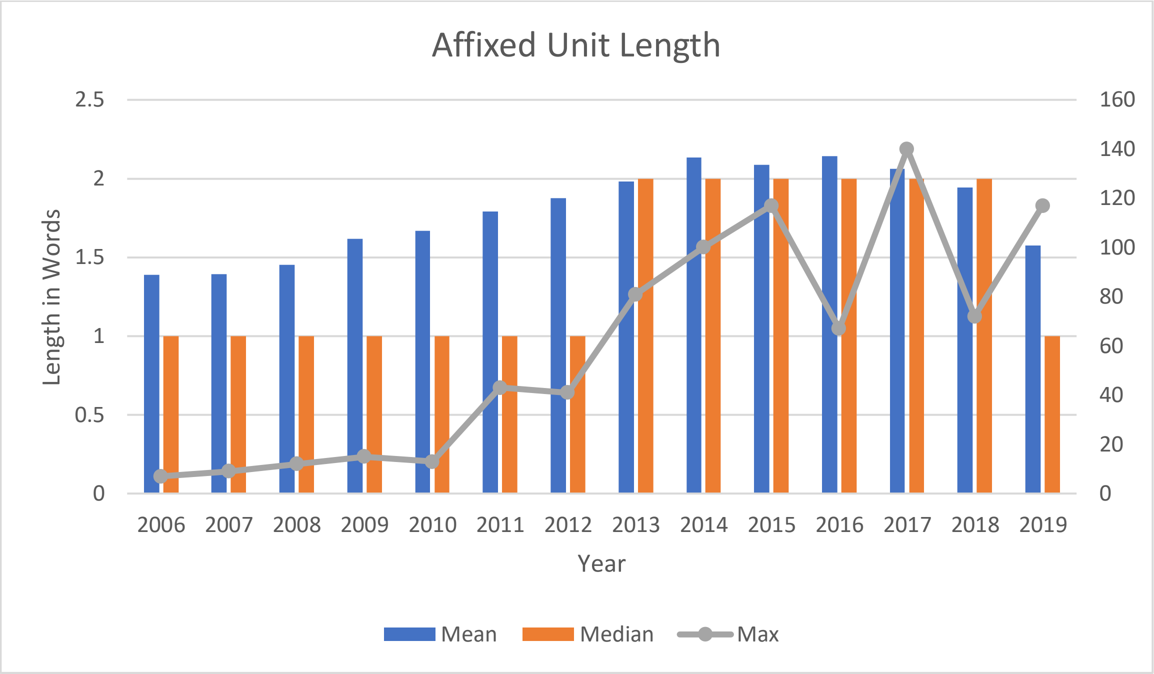
In Williams (2021) I noted that one of the unique features of AOs is their ability to affix to units with seemingly unrestricted length. That is, they have very flexible (and often ambiguous) scope in terms of the number of words that they are modifying. They may modify extremely long words of phrases as in the example 10 below.

10. I’m a lazy, stupid, cowardly person who doesn't want to have to think, so I'll just say everyone is equally bad, throw my hands up in despair, and curl up into the fetal position while mumbling about a **Complete Overhaul Of Our Completely Broken System In Which All Actors Are Equally Bad And Exactly The Same In All Respects™**
    r/politics

While there is some ambiguity in terms of the scope of the TM in the example above, the sequential capitalization leads to a possible reading in which 21 words are included in the affixed text. Affixed length was considered a helpful measure of how operator use may have changed over time, as well as understanding how frequently users take advantage of this unrestricted length.

Affixed text was extracted and segmented according to the preprocessing steps in the previous section. After segmentation, affixed unit length was generated by counting the number of words in a given extracted text. After generating affix length, high values (>50) were manually examined to ensure the automated processing had performed correctly. If an example was erroneously high, it was manually corrected, or in the case of spam-like posts (e.g., a post with hundreds of links to websites), removed. One common issue in the segmentation was that playful misspellings such as *waiiiiiiiiiiittttttttttttttttttttttttttt* were erroneously segmented into multiple words. A second regex was created to find instances like this by looking for affixed units with a character repeated >5 times in a row. This found ~12k instances which were recoded as having an affix length of 1. After this additional cleaning, affixed length was aggregated at a yearly level with mean, median, and max. These values are shown in Figure 3.6 below (non-logarithmic axis).

Figure 3.6: Affixed unit length



There appears to have been a subtle increase in affix length over time, with the mean length climbing to around 2 in 2014, before dipping back down near 1.5 in 2019. This same trend is visible by looking at the median values which increased from 1 to 2 in 2013, before dropping down to 1 in 2019. The maximum length for affixed units has substantially increased over time, with the maximum affixed units staying above 60 tokens after 2013. The longest affixed unit, at 140 tokens, is below in example 11 with the tokenized form below for readability.

11. #howlongofahashtagistoolongforcommonhashtagsbecauseicankeepaddingtomyhashtagforalongtimebecausewheniwaslittleihadahorsenamedjackwhohadanaccidentandwehadtowaitalongtimeforhimtogetbetterthenwheniwasalittleolderhegotsickandhadtostayattheveterinariansstableforalongtimeandthenwhenibecameanaduldwehadtoputjackdownsonowiwillhavetobepatientandwaituntiligettoheaventoseejackagainsoasyoucanseeivelearnedtobeapatientpersonwhichhasallowedmetowritethisreallylonghashtagisthatlongenoughtobeconsideredexcessive
r/Random_Acts_of_Amazon

> how long of a hash tag is too long for common hash tags because i can keep adding to my hash tag for a long time because when i was little i had a horse named jack who had an accident and we had to wait a long time for him to get better then when i was a little older he got sick and had to stay at the veterinarians stable for a long time and then when i became an adult we had to put jack down so now i will have to be patient and wait until i get to heaven to see jack again so as you can see ive learned to be a patient person which has allowed me to write this really long hash tag is that long enough to be considered excessive

Somewhat humorously, the contents of this hashtag are about how long a hashtag can or should be before it is considered excessive. However, even in a silly exploration of when a hashtag is ‘too long,’ the author demonstrates that according to the implicit rules that govern hashtags, this is technically a valid use.

In summary, the vast majority of AOs in the corpus affixed to single words. However, in the last 6 years of the corpus, median and mean values have seen a small shift upward. Furthermore, the emergence of extremely long affixed units in 2013 demonstrates an awareness by authors that the length of the affixed units is unrestricted, even if the majority of affixed units are relatively short in length.

3.3.3 Part of speech analysis

Analyzing part of speech information of AOs offers insight into potential changes over time, as well as potential restrictions around AO use. Furthermore, because of the origins of ©/®/™, in which the affixed unit is typically a product or entity, we might expect that the affixed units for these AOs would typically be NPs. Seeing the degree to which AOs in the corpus diverge from this expectation is an important part of building up a descriptive analysis of these operators. Accessing part of speech (POS) information involved automated POS tagging, manual annotation of tags into POS clusters, and then analysis of the POS trends.

POS tagging was done with the Python package spaCy (Honnibal & Montani, 2017). The output of this tagging was a corresponding POS tag for every word in a comment. An example affixed unit from the corpus is presented below, with its corresponding POS tags below it.

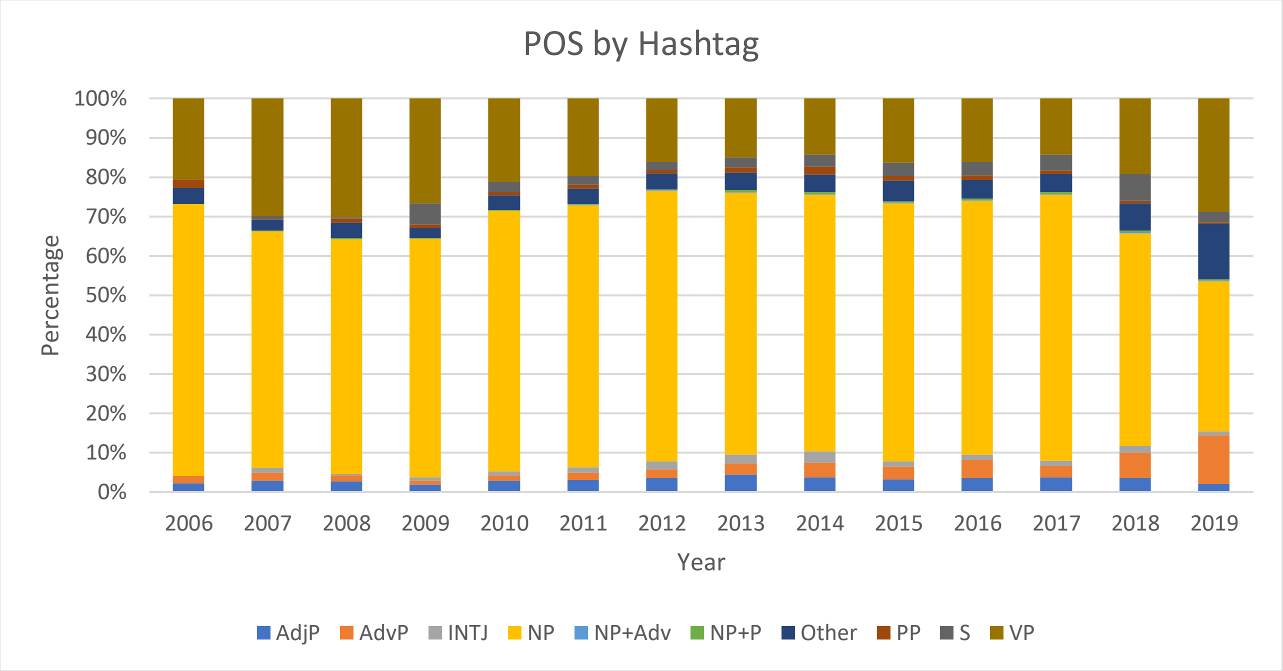
12. remove the queue
    VERB DET NOUN

Due to the size and diversity of the corpus, there were more than 100,000 unique POS tag combinations for affixed units. It was deemed implausible to hand-code these. Due to the large number (~8.4m) of AO examples, several simplifying assumptions were employed to reduce the variety of POS labels. Different combinations of labels were sampled and manually inspected, then assigned to syntactic phrase labels (e.g., VP, S). Based on the manual inspection, a set of heuristics were developed to apply this labeling to the entire corpus. If an affixed unit was only a single word, the spaCy POS tag was used, with single parts of speech (e.g., nouns) being coded as there phrasal counterparts (e.g., NPs) for simplicity. For instances where the affixed unit had 2+ words, the top 50 POS tag combinations were hand-coded into POS clusters (e.g., Verb + Det + Noun = VP). In addition to the top 50, any POS clusters that came up as a matter of course during the analysis were also manually coded. The codes used are detailed in Appendix B.

For the remaining POS tag combinations, if the first POS tag was an adjective and the last POS tag was a noun (e.g., *first world problems*), they were considered to be NPs. If the first POS tag was an adjective and the last POS tag was a verb (e.g., *black lives matter*), they were coded as sentences (S). Otherwise, the first POS tag was assumed to be the head of the phrase (e.g., Noun + Prep + Det + Noun = NP). With automating tagging, and then automatic phrase annotation, there were bound to be errors in this process, however, these methods were seen as the most practical approach for annotating the ~8.4M AOs in the dataset.

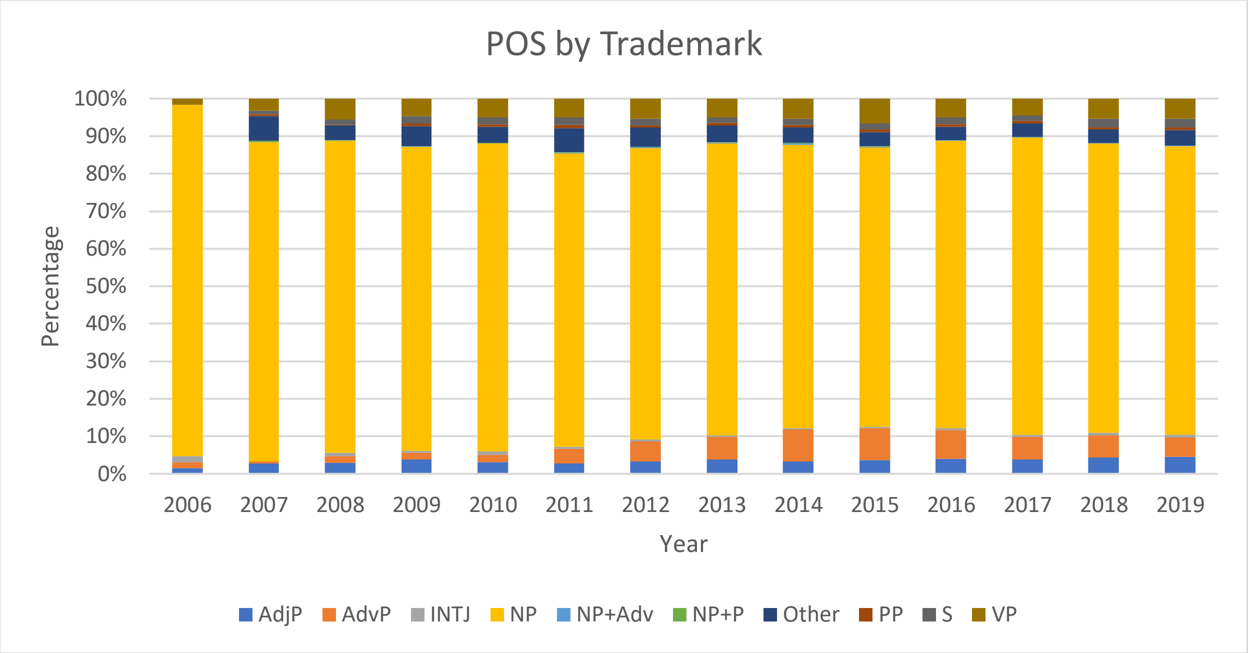
As previously stated, for the purposes of summarizing the results, I do not distinguish between individual words (e.g., a single adjective), and their phrasal counterpart. That is, I code Adjectives and Adjective Phrases both as *AdjP*. Words which were tagged as particles numbers, conjunctions, symbols, punctuation, and lone determiners were collapsed into a category: *Other*. This category also included the words that spaCy itself was unable to tag. Following this coding schema, the distribution of POS tags across the corpus for each operator is shown in Figures 3.7-10 below (non-logarithmic axes).

Figure 3.7: POS by hashtag



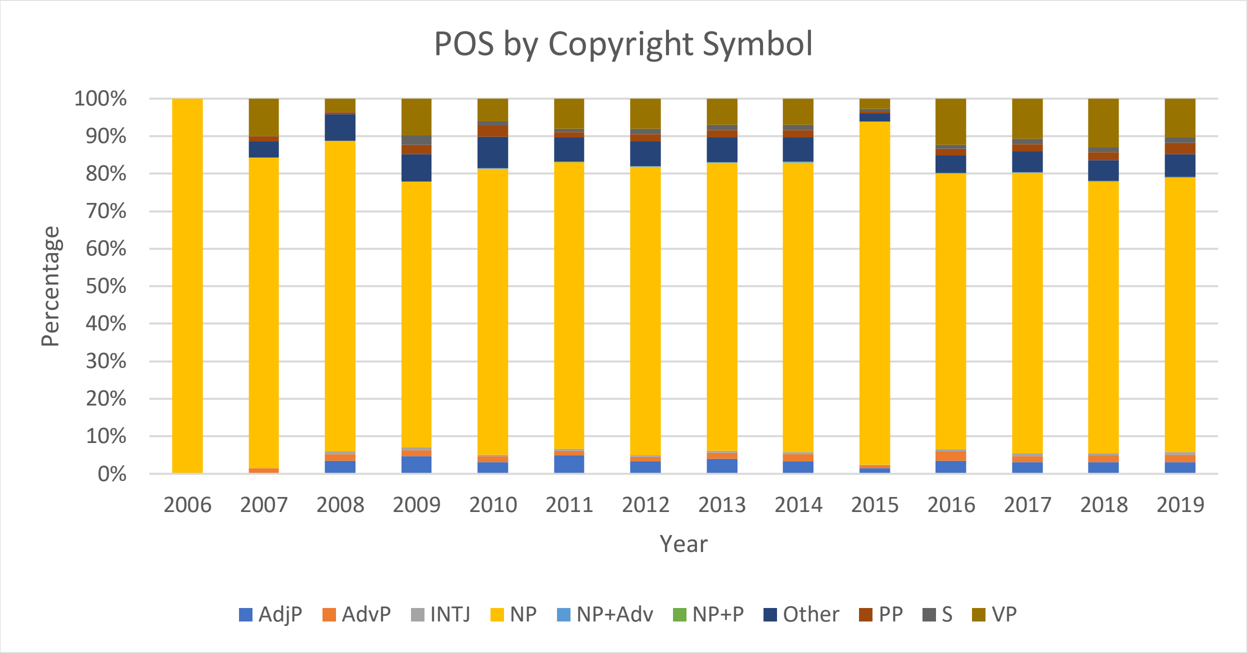
The majority of hashtags are NPs (e.g., *#everydaysexism*), although in the last two years of the corpus the percentage of NPs has decreased. VPs (e.g., *#winning*) are the second largest group. In the first six years and last two years of the corpus, VPs comprised 20-30% of hashtags. Notably, hashtags have had >14% of their affixed units be VPs across the entire corpus, which is not the case for the other AOs. Hashtags have also seen a steady increase in adverb phrases (e.g., *#reallystupid*) and NP+Adv (e.g., *#worstdadever*) in the last few years of the corpus.

Figure 3.8: POS by trademark



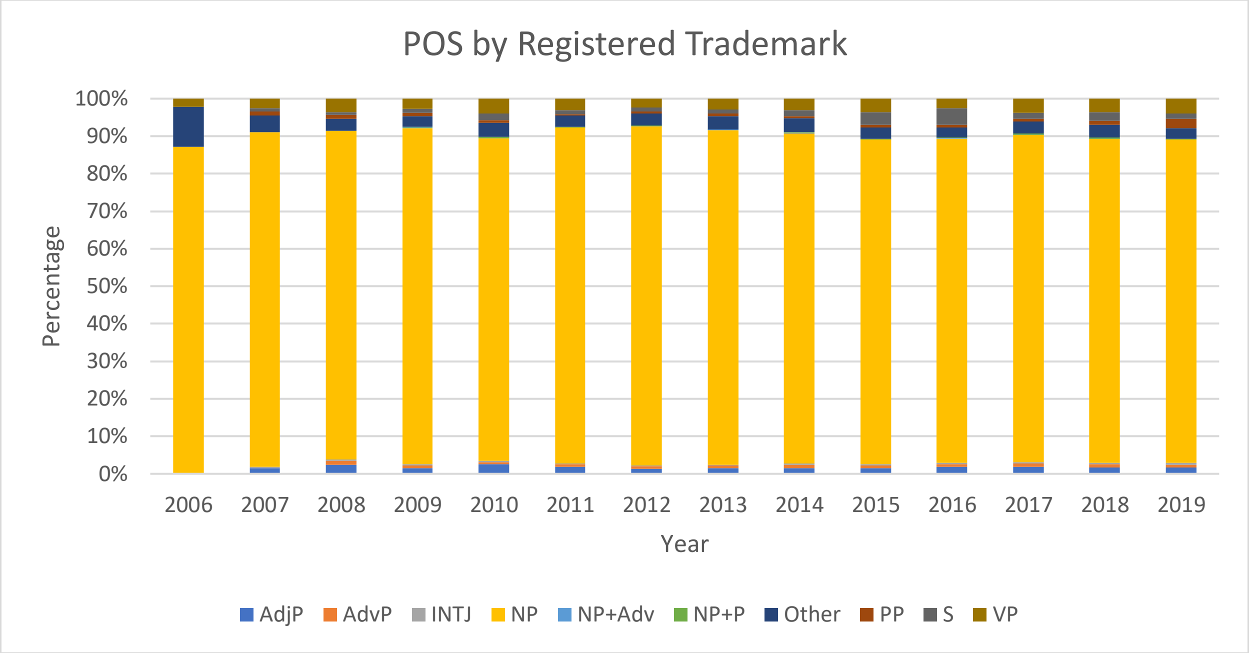
Compared with the hashtag, the trademark shows less variation over time, with the majority across the entire corpus being coded as NPs (e.g., *God-Given Conscience*™). There is additionally a slight increase in adverb phrases (e.g., *soon*™) and adjective phrases (e.g., *corporate*™) in the later years of the corpus with these phrases rising up from 3% in 2006 to comprise 8% of overall uses in 2012 and maintaining that percentage or higher for the remainder of the years in the corpus.

Figure 3.9: POS by copyright symbol



The copyright POS data is again dominated by NPs (e.g., *Same Old Bullshit*©), however, compared to the trademark has a higher overall volume of VPs (e.g., *purging*©) across the corpus, particularly in the last 4 years, with VPs constituting greater than 10%. Also notably, the copyright symbol does not see the same upward trend in adverb and adjective phrases.

Figure 3.10: POS by registered trademark



Finally, the registered trademark symbol appears to be the most stable out of all the operators, with a very consistent majority of NPs (e.g., *Mosquito Magnet*®), and a smaller percentage of other categories compared to the rest of the operators.

All of the operators show a preference for NPs. In recent years the hashtag has started to see a substantial increase in VP hashtags. The copyright and the trademark show a slight decrease in NP dominance, while the registered trademark appears unchanged. The hashtag appears to have the freest range of use, followed by the trademark and the copyright, with the registered trademark appearing to have the strongest preference for NPs.

3.4 Artificial operators and Communities of Practice

In the previous chapter, I proposed that the sociolinguistic theory of Communities of Practice (CoP) presents a useful lens through which we may examine AO use. There are three main criteria for identifying CoP: mutual engagement, jointly negotiated enterprise, and shared repertoire (Meyerhoff & Strycharz, 2013). Leuckert and Leuckert (2020) argued for subreddits as instances of CoP by demonstrating that at least in some instances, subreddits meet this criteria.

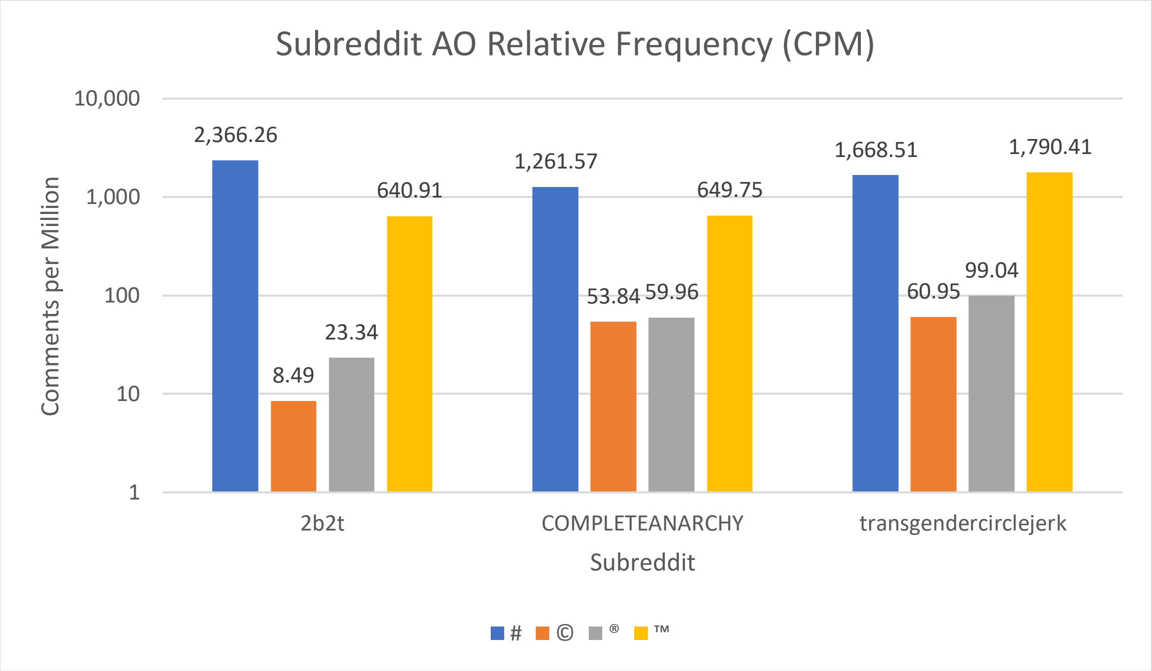
This dissertation builds on the findings of Leuckert and Leuckert (2020) and specifically analyzes the use of AOs as a component of the shared repertoire of different subreddit communities.

To perform this kind of community-based case study, three subreddits were selected from the corpus for closer analysis. For the case study, I sought to find subreddits that were thematically distinct but contained instances of all four operators. To understand the theme of the subreddit, I navigated to the subreddit home page to read a description of the community. Subreddits were carefully selected based on post volume, AO relative frequency, and topic. To ensure subreddits were large enough to have a well-established community, each subreddit selected contained at least 250,000 words. However, extremely large subreddits (more than 1M total posts) were excluded, to focus on subreddits that are more likely to have a coherent community. E.g., r/politics was excluded, since its sheer size and variety of discussions makes it unlikely to be a single coherent community (though it may host several smaller communities within it). Smaller subreddits were determined to be more likely to host a single community, and thus, a better target for this analysis. Finally, AO relative frequency was examined, to ensure that all three communities had a regular pattern of AO usage.

The three subreddits selected were transgendercirclejerk, 2b2t, and COMPLETEANARCHY. Subreddits were chosen such that they each have a distinct topic that is specialized enough to create a sense of community, without expecting tremendous overlap between the communities. Transgendercirclejerk is described as a "a parody subreddit for trans people, mocking all transgender-related topics" (R/transgendercirclejerk, 2022). 2b2t (2builders2tools) is a subreddit dedicated to a public server for the game Minecraft.

COMPLETEANARCHY is a community for political anarchists to share memes. The relative frequency of AOs in comments per million (CPM) is below in Figure 3.11.

Figure 3.11: Case study subreddit AO relative frequency

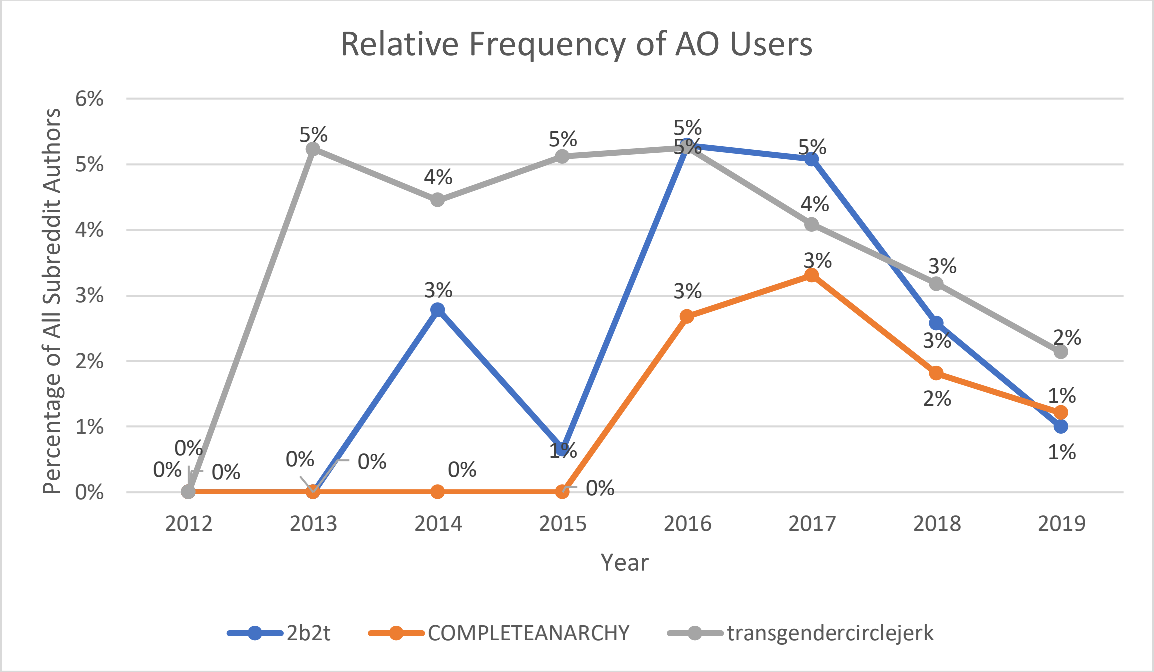


2b2t and COMPLETEANARCHY had the most hashtags, followed by trademarks. Transgendercirclejerk, notably, had more trademarks than hashtags. However, these two operators were more frequent than © and ® across all three subreddits. In the following sections, I investigate these phenomena more extensively through a CoP framework. In section 3.4.1, I look at the top users of AOs within each community and compare their AO usage within the case study subreddits to their AO usage in other subreddits. In section 3.4.2, I examine the degree to which participation in the case study subreddits may be indicative of membership within a broader community. In section 3.4.3, I explore the top AOs for each subreddit, to understand the degree to which AOs may be a valuable example of shared repertoire in CoP. In section 3.4.4, I look at how often affixed units were repeated in each subreddit, as opposed to novel pairings between affixed units and AOs. In section 3.4.5, I compare the affixed unit length and the POS cluster distribution of the case study subreddits to the broader corpus. Finally, in section 3.4.6, I look at one instance from a case study subreddit where an affixed unit was used with all four operators.

3.4.1 User behavior

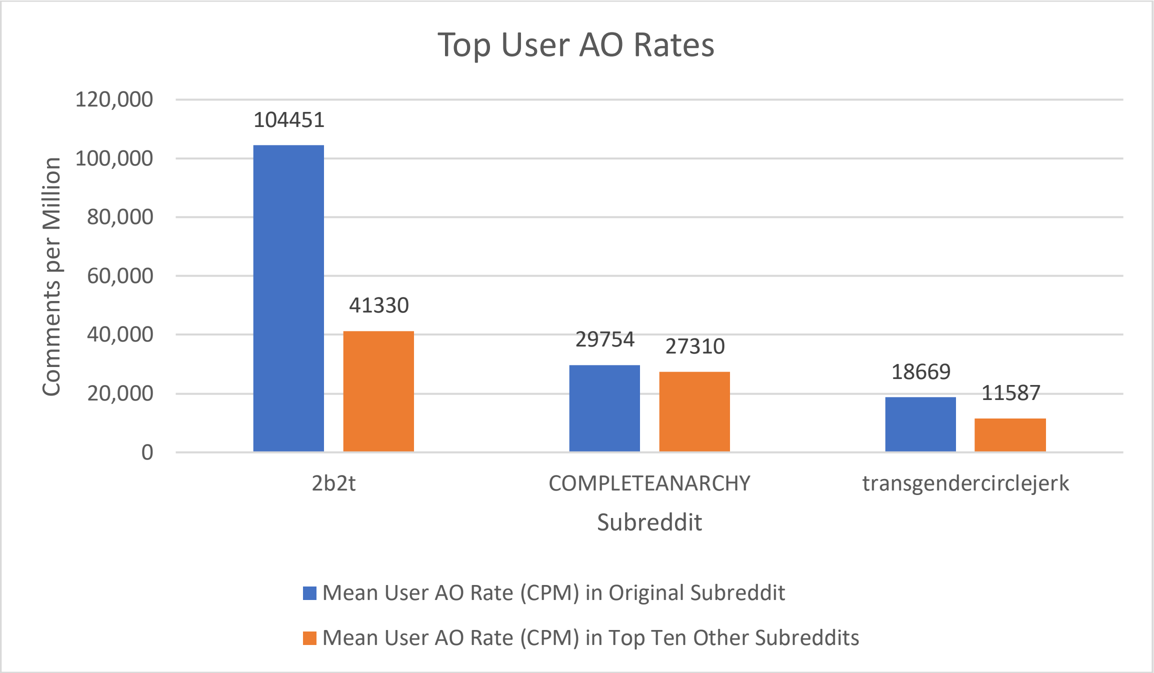
With the concept of Communities of Practice in mind, the first thing that I looked at was user behavior. Figure 3.12 (non-logarithmic axis) shows the yearly percent of authors within each subreddit who used any of the four AOs.

Figure 3.12: Relative frequency of AO users in case study subreddits



In each of the three subreddits, less than 10% of authors use AOs. As the focus here is on AO-usage as a potential manifestation of CoP shared repertoire, I focused on high frequency AO-users, pulling the top ten authors by AO comment volume from each subreddit. For each of those 10 authors, I looked at their comment history across the full corpus and identified their top 10 subreddits (or all subreddits, if the user did not post in at least 10), excluding the case study subreddit they were already associated with. This yielded 281 non-case study subreddits. It should be noted that one of the top users in 2b2t did not participate in any other subreddits.

Figure 3.13: Case study top user AO rates



In Figure 3.13 above (non-logarithmic axis), I have aggregated the 30 top AO-users (10 per subreddit) and calculated two different mean AO rates. In each of the case study subreddits, authors were more likely to use AOs in the case study community compared to their non-case study subreddits. I make two observations about this finding. The first is that AO use is not restricted to one subreddit – instead, authors who use AOs in one community are likely to use them in another. The second is that these authors, on average, used AOs at a higher rate within the case study communities. This provides evidence that there is a community-level effect on the authors' use of AOs, and that authors modify their use within the different CoP in which they participate.

3.4.2 Degrees of community

In addition to looking at user behavior in terms of AO frequency, I examined the non-case study subreddits for potential thematic overlap with the corresponding case study subreddit. Each of the 281 non-case study subreddits was given a similarity score in relation to the original subreddit. In order to assign this score, I navigated to the subreddit home page and read the description to ascertain the topic of the subreddit. I then determined if the topic was closely related, loosely related, or unrelated. This similarity framework is repurposed from Leuckert & Leuckert (2020). The rating system, with the topics I looked for, is below in Table 3.1.

Table 3.1: Rating system used for thematic cohesion scores

	2b2t	**transgendercirclejerk**	**COMPLETEANARCHY**
3 (extremely closely linked to the topic)	Minecraft; Minecraft servers, Minecraft community	Transgender topics	Anarchism
2 (loosely related to the topic)	Video games	LGBT; Gender	Other political ideologies; Politics; Society
1 (unrelated or unclear)	any other topic	any other topic	any other topic

For the subreddit COMPLETEANARCHY, if the description was explicitly related to the same topic (e.g., anarchism), it was rated as a 3. If the description was loosely related to the topic (e.g., a political ideology that is not anarchism), it was rated as a 2. Otherwise, it was rated as a 1. This process was repeated for each subreddit using the topics in Table 3.1.

After assigning each of the 281 non-case study subreddits a numeric rating, thematic cohesion scores were generated for each case study subreddit. Cohesion scores for a subreddit were calculated as the mean similarity between that subreddit and the non-case study subreddits. Higher scores indicate that the sampled authors participated more in subreddits related to the topic of the case study, while lower scores indicate that the sampled authors participated in a broader range of topics. Subreddits could occur more than once in the calculation if more than one author had that subreddit in their top ten. This was done because if every single author in the sample were to participate in the same closely related subreddit, this is a powerful indicator of cohesion that would be lost if repeated subreddits were only counted once.

Figure 3.14: Thematic cohesion scores for each case study subreddit

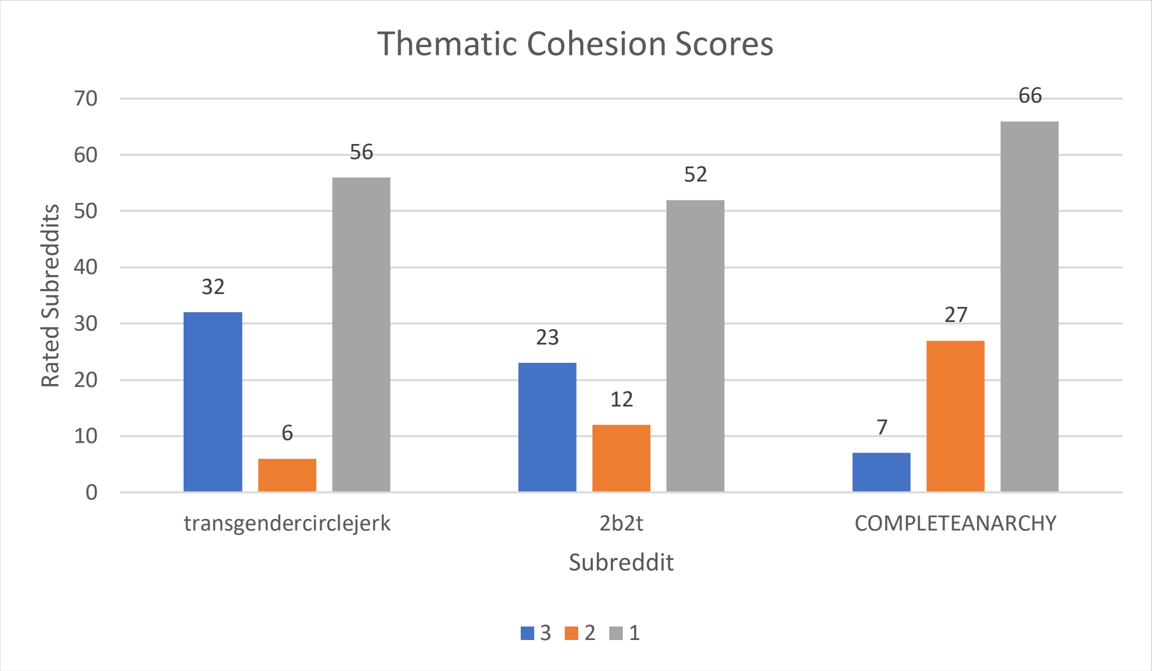


Figure 3.14 above (non-logarithmic axis) shows the distribution of scores for each of the case study subreddits. Transgendercirclejerk had the highest thematic cohesion score with a 1.74. Of the rated subreddits for transgendercirclejerk, 32 were scored as a 3, meaning they dealt explicitly with transgender topics. Notably, only six of the rated subreddits received a 2, suggesting that participation in transgendercirclejerk did not necessarily indicate participation in other communities dealing with LGBT or gender issues. 2b2t had a thematic cohesion score of 1.67. Users from this community were likely to participate in in other communities related to Minecraft (23 ratings), and communities related to video games more broadly (12 ratings). COMPLETEANARCHY had the lowest cohesion score with 1.41. However, it is interesting to note that while there were fewer rated subreddits which were explicitly related to anarchy (7 ratings), they had the highest volume of rated subreddits which received a 2 (27 ratings). High AO usage in COMPLETEANARCHY was a good indicator of participation in other communities dealing with politics and political ideologies, but not necessarily other communities

dealing specifically with anarchy. In other words, being a high AO user in the transgendercirclejerk and 2b2t subreddits seems to be a reliable indicator for membership in the broader transgender and Minecraft communities respectively. For COMPLETEANARCHY, being a high AO user did not show as strong of a relationship of being part of the broader anarchist community but did seem to indicate membership in communities dealing with politics. Interestingly, 2b2t had the highest relative frequency of AOs for both the case study and non-case study subreddits. In other words, 2b2t shows the strongest evidence for AO-use to be a sign of community membership potentially within the broader Minecraft community.

3.4.3 Shared repertoire

The case study allowed for me to look not just at broad quantitative trends, but also at the specific affixed units that are being produced within these communities. For each of the three case study subreddits, I generated the five most frequent affixed units for each operator. AOs were examined based on their raw frequency as well as their contextual diversity (see Baker & Subtirelu, 2017). Contextual diversity (CD) is an indicator of how widely dispersed a word is throughout a corpus, and it can mitigate scenarios in which a small portion of the corpus might be over-represented. It is often calculated as the percentage of texts within a corpus which contained a given word. For the purpose of calculating CD, I considered each complete year that the subreddit existed to be a separate text. For example, 2b2t and transgendercirclejerk were founded in 2012 and COMPLETEANARCHY was founded in 2015. In calculating CD, for 2b2t and transgendercirclejerk I treated every year from 2013-2019 as a text and for COMPLETEANARCHY I treated every year from 2016-2019 as a text. Affixed units were transformed into lower case to avoid double counting those that differed only by capitalization. For each of the top affixed units, examples were manually examined to ensure that the text had been correctly extracted. Posts which were created by bots were removed where identifiable. In cases where it was clear from context that the automatic extraction had resulted in a fragment of the overall affixed text, manual corrections were made. (e.g., tourism – 2b2t Board of Tourism). Affixed units which only occurred once in the subreddit were not included in the results, as there were cases where more than 40 affixed units had a frequency of one, and it did not make sense to include some of these without including them all. The results for the 2b2t subreddit are below in Table 3.2.

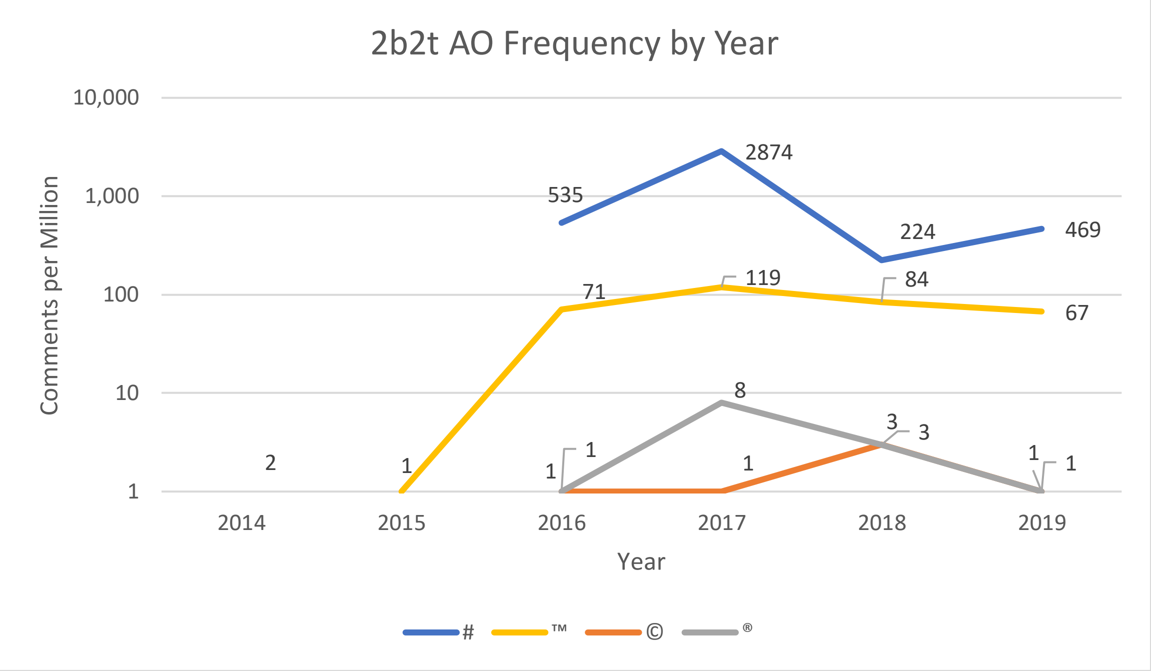
Table 3.2: Top AOs in 2b2t

2b2t Top AOs			
Operator	**Affixed Text**	**Volume**	**CD**
#	teamvetrain	2631	0.57
#	savemaps	103	0.14
#	teamveteran	91	0.57
#	teamrusher	78	0.57
#	the	37	0.57
™	tourism	59	0.29
™	soon	28	0.57
™	team uberslugcake	11	0.14
™	minecraft	8	0.43
™	team tortellini	7	0.14
©	ghast	2	0.14
®	intel	2	0.29
®	pepsi products	2	0.14
®	voco	2	0.14

A number of the affixes in 2b2t are some variation of team ____, referring to people showing their alignment with various factions within the community. The copyright symbol and registered trademark symbol are very uncommon within this community, as such there was only one repeated affixed unit for the copyright symbol and only three for the registered trademark symbol. Hashtags had the highest CD and the highest frequency, but on the whole, none of the

AOs have a particularly high CD in 2b2t. I additionally aggregated the volume of AOs for 2b2t for each complete year of the subreddit. This is below in Figure 3.15.

Figure 3.15: 2b2t AO frequency by year



The frequencies appear to be in line with overall trends of the corpus, with hashtag the most popular, followed by the trademark, and then copyright and registered trademark.

The top AOs for the subreddit COMPLETEANARCHY are below in Table 3.3.

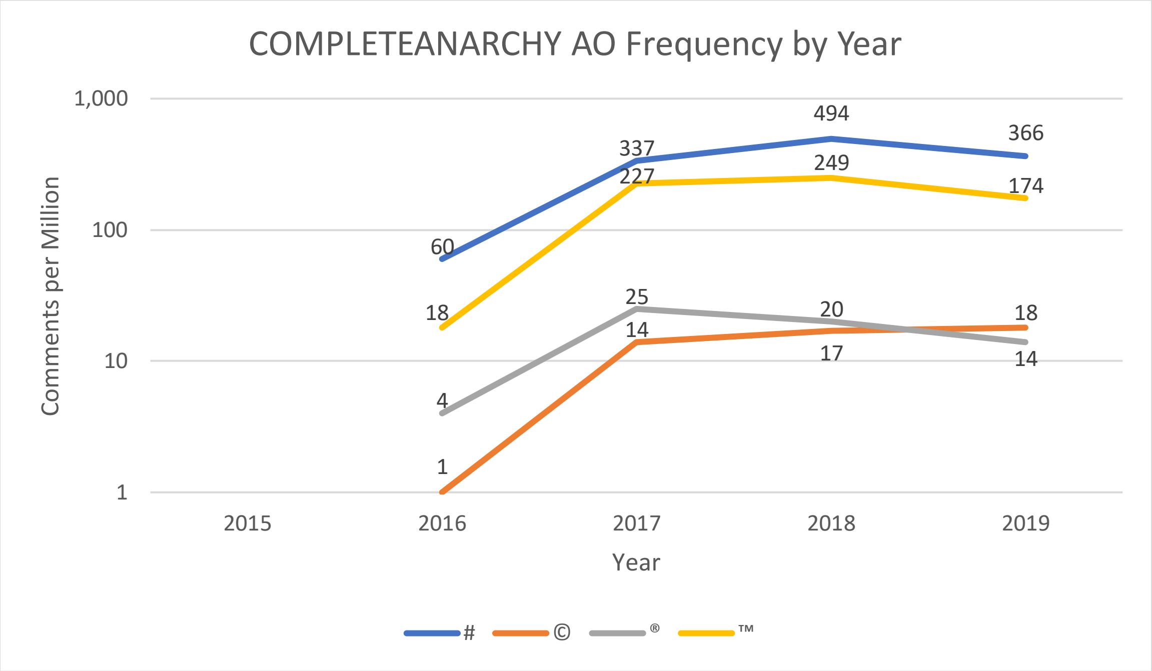
Table 3.3: Top AOs in COMPLETEANARCY

COMPLETEANARCHY Top AOs			
Operator	**Affixed Text**	**Volume**	**CD**
#	resistance	32	1.00
#	metoo	30	0.50
#	the	26	0.75
#	wegotthis	22	0.50

#	chapter	16	0.25
™	freedom	10	0.50
™	market	8	1.00
™	nap	8	0.75
™	free market	7	1.00
™	mcpolice	6	0.75
©	permission	3	0.25
©	green capitalism	3	0.25
©	mcdonalds	2	0.25
©	herrenrasse	2	0.25
®	police	5	1.00
®	greek yogurt presents handcuffs	3	0.75
®	freeze scumbag	3	0.75

Compared to 2b2t, where top AOs dealt with gaming terms or specialized in-game language, AOs in this community have thematic threads with political and societal issues. There are references to political movements like the #MeToo and #Resistance movements, as well as economic and political buzz words such as *freedom*, *free market*, and *capitalism*. Compared to 2b2t, COMPLETEANARCHY had higher overall CD, suggesting that there are popular AOs which are used year after year within this community. On the other hand, even for hashtags, COMPLETEANARCHY did not have any AOs with extremely high frequency, such as #teamvetrain in 2b2t. While there are common phrases that are used across multiple years, it is not the case that a handful of popular affixed units have driven up relative frequency of the subreddit. Instead, it would appear that operators with a broader range of affixed texts has made this community have a high relative frequency. The frequencies for AOs in COMPLETEANARCHY across the years are below in Figure 3.16.

Figure 3.16: COMPLETEANARCHY AO frequency by year



Unlike 2b2t, COMPLETEANARCHY appears to have some minimal usage of © and ® across all years. Additionally, the gap between the trademark and the hashtag is less substantial here.

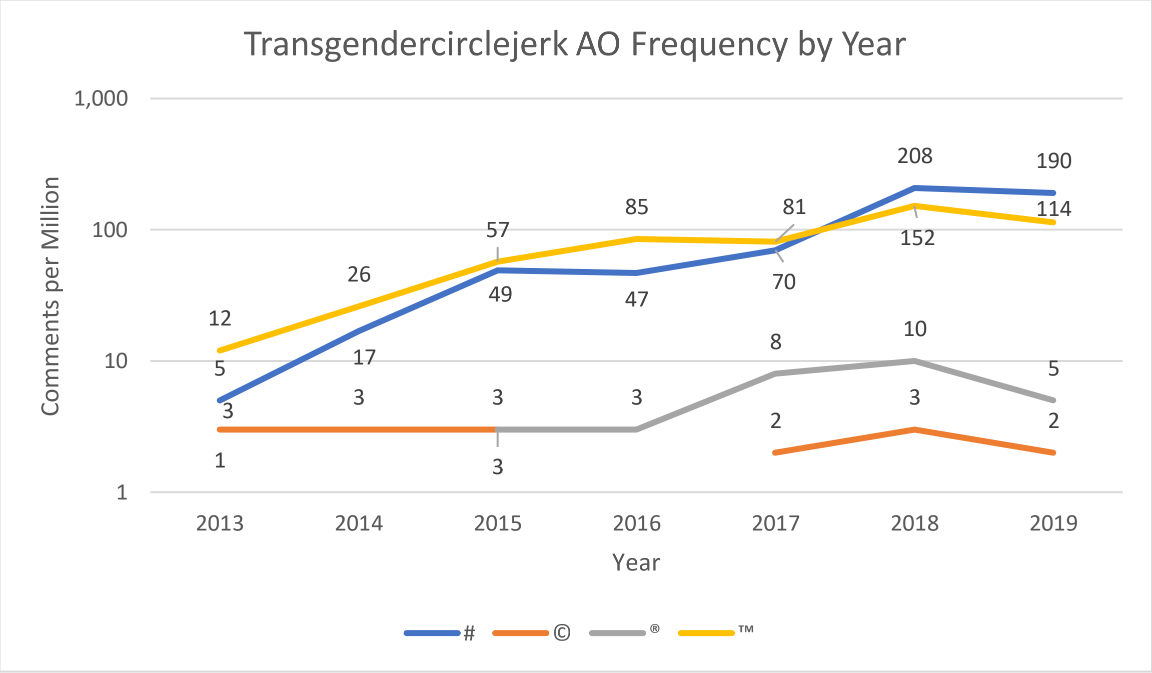
Lastly, the top affixed units for the subreddit transgendercirclejerk are below in Table 3.4.

Table 3.4: Top AOs in transgendercirclejerk

transgendercirclejerk Top AOs			
Operator	**Affixed Text**	**Volume**	**CD**
#	goals	12	0.29
#	metoo	7	0.43
#	the	5	0.43
#	maga	5	0.43
#	you	5	0.29
™	trutrans	133	1.00
™	trans	12	0.71
™	transgenda	10	0.43
™	truetrans	7	0.71
™	woman	7	0.71
©	trutrans	2	0.29
®	trutrans	5	0.29

Transgendercirclejerk is notable in being the only of the three communities where the most popular affixed unit is a trademark, not a hashtag. While *Trutrans™* has a CD of 1, indicating it is used in every year of the subreddit corpus, many of the AOs in this list have low CD. The copyright and registered trademark each had only one repeated affix: *trutrans*. Looking at the content of the affixed units, many are related to transgender issues or political movements (e.g., *metoo*). As this is a parody subreddit, some of the affixed units imitate transphobic rhetoric, such as #*maga* or being *trutrans*. The frequencies by operator across time for this subreddit are below in Figure 3.17.

Figure 3.17: Transgendercirclejerk AO frequency by year



Again, this subreddit stands out for having very high use of the trademark, that unlike the broader corpus, actually surpasses the hashtag in 2013-2017.

All 3 subreddits are similar in that the use of the © and ® is relatively rare. Additionally, some of the extracted AOs from these subreddits show authors using the symbols in ways counter to my expectations or the patterns my AO-extraction regex was looking for. For example, authors leveraging the symbol to replace a letter (e.g., ®usher) or copyrighting comments with their username (e.g., © RightKnight27). However, the affixed units are overwhelmingly related to the topics of the community at hand. In other words, AOs in these 3 subreddits are most frequently used with highly community-specific terms.

There were 29 AOs that occurred across all 3 subreddits. All of them were hashtags except for one: soon™. I do a close examination of the *soon™* AO in chapter 5, due to its prevalence across many communities. Other shared hashtags between the communities included things like *#relatable* and *#owned*.

3.4.4 Repeated affixes

Each of the three subreddits was assigned a score for how linguistically productive the operator was. This was calculated as the number of unique affixed texts divided by the total number of uses of the operator. In other words, this was a measure of how likely the operator was to be used with new affixes versus repeatedly being used with the same texts. These scores are in Table 3.5 below for each of the operators.

Table 3.5: Repeated affix scores for case study subreddits

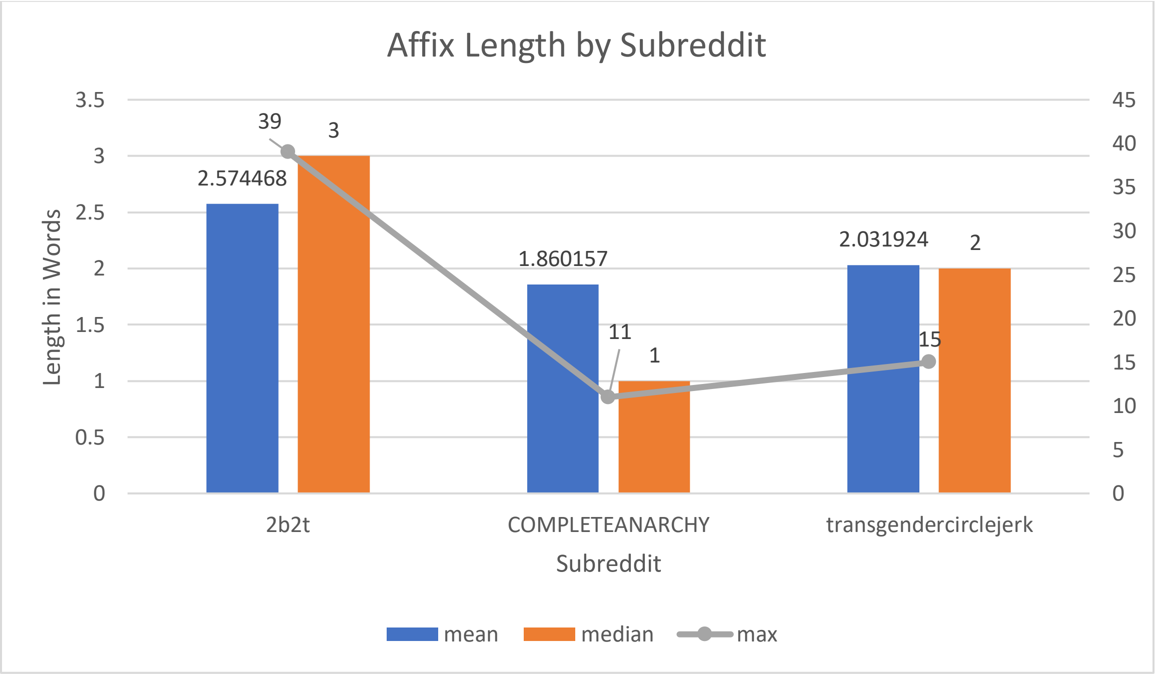
	Transgendercirclejerk	**2b2t**	**COMPLETEANARCHY**
#	0.76	0.16	0.67
®	0.87	0.77	0.87
©	0.94	0.83	0.88
™	0.51	0.54	0.76
MEAN	0.76	0.57	0.75

2b2t had the lowest score for the hashtag with 0.16, suggesting that the hashtag has overwhelmingly been used with the same affixes (e.g., #TeamVetrain). Interestingly, transgendercirclejerk had a fairly high score for the hashtag, suggesting that it is often used with novel affixed units. By contrast, COMPLETEANARCHY had the highest score for the trademark, suggesting that it is often used with novel affixed units. There were very high scores for the ® and © across the board, which is likely related to the fact that the operators were used in such low frequencies.

3.4.5 Comparing the case study dataset to the AO dataset

In order to understand the degree to which these subreddits might differ from the broader corpus, affixed unit length was calculated. These are below in Figure 3.18 (non-logarithmic axes).

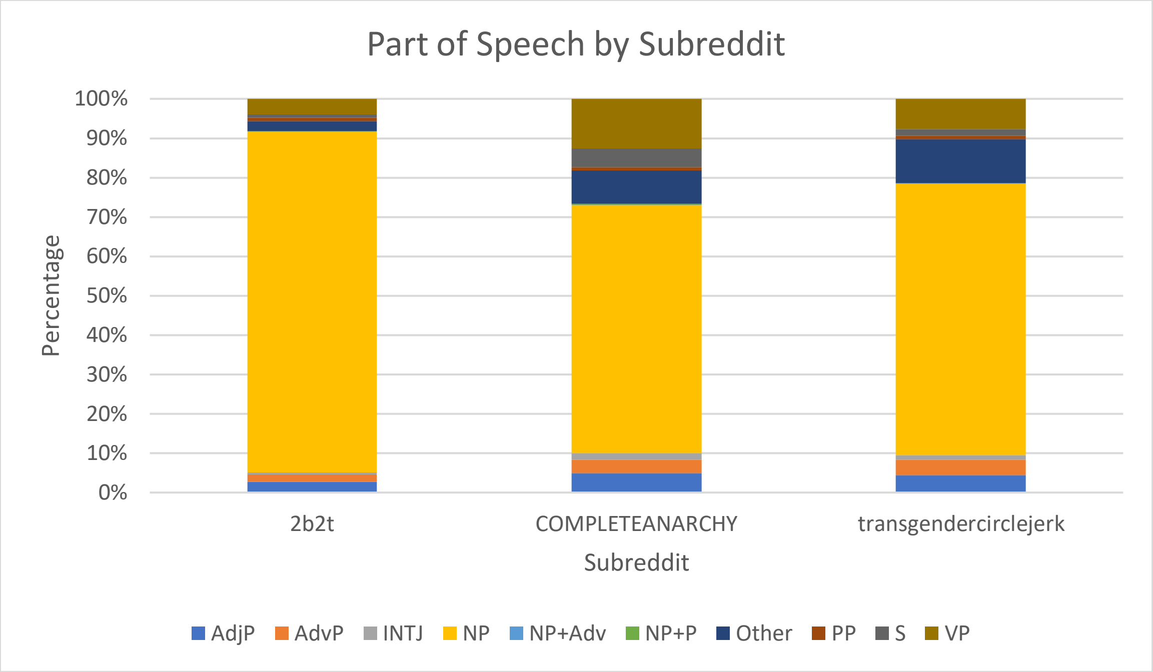
Figure 3.18: Case study affixed unit length by subreddit



These subreddits were selected for having a high relative frequency of AOs. Interestingly, each of the three subreddits has a higher mean and median affix length compared to the full corpus (Figure 3.6). This could suggest a relationship between being a frequent user and affixing the operators to more text.

Part of speech tagging was also performed on the case study subreddits. The results are in Figure 3.19 below (non-logarithmic axis).

Figure 3.19: Case study part of speech distribution by subreddit



As was the case with the full corpus, NPs were the largest group across all 3 subreddits. None of the subreddits differed substantially from the full corpus with respect to their POS distribution. However, 2b2t had a much larger percentage of NPs compared to the other two subreddits, likely due to the prevalence of the *#team* ____ hashtags. It is interesting to note that the two subreddits with a more varied distribution of POS clusters also had higher overall productivity scores. In those communities where AOs are used more productively, in other words, they occur not only with a broader range of lexical items but also in a freer range of POS contexts.

3.4.6 Interchangeability

The last area I examined in the case study analysis was an exploration of the potential interchangeability of the AOs. In order to approximate this interchangeability, I searched for affixed units which co-occurred with all four AOs. The case study presented an opportunity to more closely examine these instances and determine from context whether they were being used to similar effect. Across the case study corpora, there was only one affixed unit that was shared across all 4 operators. This was *TruTrans* from the subreddit transgendercirclejerk. Examples with each of the operators are below in examples 13-17.

13. Nah, but I will have to revoke your **#TruTrans** card

14. I never thought of that! You're right though, maybe I can be **TruTrans™** after all!

15. I'm already an awful driver. Should I detransition for the safety of other drivers or does it just mean that I'm **TruTrans®**

16. I'm not on spironlactone. I suck souls out through men's semen so they go gay to maintain my womynlyness. I'm **trutrans©** bitch.

17. Make sure you also knock over as many butter dishes as you can with your girl dick, so people will know you're **TruTrans©®™**

The examples above show authors using all four AOs to mark the slang, community-specific term *TruTrans*. I previously established that this is a parody subreddit community, where authors imitate and mock transphobic concepts. In the examples above, the AOs likely contribute to this parody by calling attention to the fact that *TruTrans* is not a real term or a real designation. There appears to be a distancing effect, where the author is signaling their own rejection of the concept *TruTrans*. I talk extensively about this distancing function in Chapter 4. However, here I would like to observe that in the above contexts, the operators do appear to serve similar effects when affixed to the same text. While authors demonstrated a preference for affixing *TruTrans* with ™ (133 instances), authors also use these alternate linguistic resources to achieve the same effect.

3.5 Conclusion

In this chapter I have carried out analysis using a Corpus Linguistics framework using datasets extracted from the Pushshift Reddit dataset. I have leveraged the entire corpus to obtain frequency counts. I have created an AO dataset of all the comments containing AOs to perform affixed unit extraction and POS analysis. Finally, I performed a case study on three subreddits using a dataset of all comments and AOs from those communities.

The frequency analysis revealed that of the operators, only the hashtag is increasing in terms of relative frequency. The trademark, copyright and registered trademark continue to be highly specialized in nature. However, it is worth noting that because I look at all instances of the hashtag here, I have not filtered out instances where the hashtag may be used for unconventional purpose (e.g., as a bullet point) or those more traditional instances where the hashtag is used for meta-evaluative commentary. Therefore, the frequencies reported here should be assumed to be an over-estimation, though verifying this and quantifying the degree of over-estimation will be left to future research and refinements.

I examined the affixed unit length across the entire corpus. This analysis revealed that the affixed unit tends to be one word, however there is a slight lengthening in the corpus where the medians get higher over time. I also showed that the length is unrestricted and users, particularly in recent years, sometimes capitalize on this for a humorous effect. This lack of restriction on the length of affixed units provides strong evidence that these operators have become part of the productive language use in communities online. However, the tendency for short lengths indicates the presence of strong norms around the use of these operators.

In terms of part of speech clusters, I showed that the operators have a strong preference for NPs. I showed that hashtags have the most variability in terms of POS cluster, with a sizable portion of hashtags affixing to VPs. Furthermore, the analysis revealed changes over time for three out of the four operators, with the operators showing some degree of decrease in the overall percentage of NPs.

The case study supported previous findings that subreddits are indicators of community membership and present instances of CoP in CMC. The case study also revealed that AO usage as a practice varies by community, even among authors who are high frequency users. Thematic cohesion scores showed that for two of the three communities, being a high frequency AO user was a reliable indicator of being a member of a broader community.

The case study subreddits did not vary substantially in terms of their POS distribution from the AO corpus, but they did have higher means and medians for affixed unit length, suggesting a possible relationship between being a high frequency user and having longer affixed unit length.

One of the trends that has begun to emerge in this corpus analysis is a difference between the hashtag and the other AOs. It appears clear that in terms of frequency and POS cluster distribution, not to mention being a prefix vs. a suffix, the hashtag is generally used in different ways. I therefore do not conclude that the hashtag has no differences; it definitively has other applications that are different from the rest of the operators. However, one of the important findings of the case study was that these symbols share properties such that they are capable of performing similar functions in the same community and being swapped out to achieve similar effects.

The case study also demonstrated that AOs are valuable examples of shared repertoire in CoP, with nearly all the common affixed units being thematically related to the community topic, and often requiring complex understanding of the community purpose to accurately interpret the meaning. For example, *TruTrans* is the most common affixed unit in the community transgendercirclejerk and co-occurs with all of the AOs. However, understanding what is meant by *TruTrans* requires understanding the purpose of the community, the ideologies of the community members, and the people with whom the community is aligned or dis-aligned. In other words, AOs are often complex expressions of interpersonal meaning that are linked to identity and positioning. As such, in the wake of the corpus analysis, the stance theory was determined to be a useful theory for performing closer analysis of AOs. I explore AOs through the lens of stance in the following chapter.

CHAPTER 4

ARTIFICIAL OPERATORS AS STANCE MARKERS

In chapter 3, I provided a descriptive analysis of artificial operators (AOs) in the Pushshift Reddit dataset. I concluded by proposing that interpersonal meaning and ideology appear to play a significant role in AO use. As such, the linguistic theory of stance provides a useful framework for studying the meaning that AOs contribute. In this chapter, I provide an analysis of AOs as stance markers. In section 4.1, I establish AOs as sharing metadiscursive properties which motivate my framing them as contextualization cues. In section 4.2, I provide background on the linguistic theory of stance and the core components of evaluation, positioning, and alignment. In section 4.3, I propose a typology of AOs based on the two speaker-oriented dimensions of stance: evaluation and positioning. In section 4.4, I look at the evaluative meaning contributed by AOs. In section 4.5, I explore how AOs contribute to speaker positioning. In section 4.6, I discuss the role that AOs play in signaling alignment with an audience. Finally, in section 4.7, I summarize the findings and propose conclusions.

4.1 Background

In Chapter 1, I outlined qualities which AOs share that motivates their study together as a group. In Chapter 3, I suggested that interpersonal, contextual meaning is a key component in understanding AOs. In this section, I propose that a fundamental feature of AOs is that they are metadiscursive in nature. I adopt the term metadiscourse following Zappavigna (2018, p. 36) to "refer to the whole gamut of interpersonal resources available for managing how a text positions itself in relation to its real or potential audience." Zappavigna uses metadiscourse to explain properties of the hashtag, namely the way that hashtags create two orders of meaning: tagged text and untagged text. These two orders of meaning are illustrated by the example 18, adapted from Zappavigna (2018, p. 30) with the tagged text in bold and the untagged text underlined.

18. **#ManyPeopleAreSaying** the four person hair-care team that failed to cover the hairpin when Trump was in Mexico have been replaced for debate

Zappavigna discussed this special information status as a function of hashtags, however, Williams (2021) proposed that this function is extensible to other AOs via examples 19-21 below (adapted from Williams, 2021, p. 166).

19. I'm sure **Brave Patriots™** will be lining up to condemn this disrespectful, anti-American violation of the Flag Code.

20. Unfortunately, plans **changed©** and it turns out I'll be in the bus for most of the show … But thank you for your answer ! Enjoy Mania !

21. That's just god trying to show you the **Truth®**

In other words, a fundamental feature of these AOs, regardless of whether they are being used figuratively, is assigning special information status to text within the sentence. It is this metadiscursive quality that leads to these operators being used to achieve similar stance-marking functions, as I will demonstrate later in this chapter. However, as I am interested in the figurative and communicative purposes of these operators, I focus on those instances where the special information status is used for a pragmatic effect, to convey speaker meaning and orientation. In other words, I am interested in situations where AOs act as ‘contextualization cues’ (Gumperz, 1992).

4.1.1 Contextualization cues

A contextualization cue is a linguistic sign which does not have meaning without context (Auer, 1992). These cues are used

> to relate what is said at any one time and in any one place to knowledge acquired through past experience, in order to retrieve the presuppositions they must rely on to maintain conversational involvement and assess what is intended (Gumperz, 1992, p. 230).

In order to interpret the meaning of a contextualization cue, the reader must rely on ‘contingent inferences’ about the context. Cirillo (2019, p.9), demonstrates the ways that air quotes act as contextualization cues, and argues that air quotes “have no conventionalized meaning, but can only be understood in relation to the context in which they are situated.” Due to the multifunctionality of AOs, and the importance of background information to interpret their meaning, AOs may similarly be characterized as contextualization cues. Other than the conventionalized function of assigning special information status to text, the interpretation of AO meaning relies on inferences about the author, intended audience, and community.

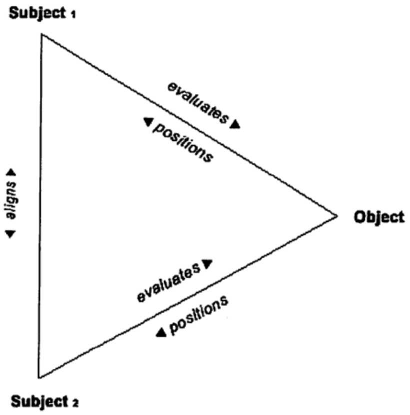
In conceiving of AOs first as contextualization cues, and then as stance markers, I aim to capture the nature of the special information status that they assign the text. The AO segments the affixed text out from the rest of the sentence, and then signals to the reader that for this particular text they need to activate context and background information in order to interpret the meaning. As for what contextual information the reader needs to access, that varies greatly. However, it is clear that at least one function of AOs as a contextualization cue is the expression of evaluative or attitudinal information, or expressions of stance. In the rest of this section, I will analyze AOs as stance markers. In the case of stance marking, the context that needs to be activated is information about the author, the audience, the subreddit community, and the typical opinions or experiences therein. To put it simply, the role of the AO is to disambiguate potential subtle contextual meaning related to the way the author orients themselves and the reader to the affixed text. In the next section, I will review previous approaches to stance and explain how I intend to use stance marking as an analytical framework.

4.2 Stance

In this chapter, I posit that a core function of AOs is to act as stance markers. For the purposes of this analysis, I adopt the definition of stance as “the lexical and grammatical expression of attitudes, feelings, judgements, or commitment” (Biber & Finegan 1989, p. 124). Stance has been labeled with various terminology across subdisciplines in linguistics such as evaluation (Hunston & Thompson, 2001), metadiscourse (Kopple, 1985), or systemic functional linguistics’ appraisal (Martin & White, 2003). However, each of these different approaches fundamentally study how language is used to create interpersonal meaning and signal attitudes and ‘stances’ toward other entities, whether objects or interlocutors. Crucial to the approach that I adopt here, stance involves the evaluation of an object, and the positioning of the author in relation to that object.

Du Bois (2007) presented a framework for analyzing stance that distills stancetaking into three components: evaluation, positioning and alignment. This is demonstrated in Figure 4.1 below.

Figure 4.1: Stance Triangle adapted from Du Bois (2007)



In Figure 4.1, Subject 1 is the stancetaker, or in the context of Reddit posts, the author. Subject 2 is the second interlocutor, or the intended audience of the post. Crucially, in a social media context like Reddit, the audience may be a specific individual, a hypothetical individual, or an entire community. The edge of the triangle that connects the two Subjects is alignment. During a stancetaking event, the author expresses a degree of alignment with the intended audience. This alignment may be convergent, in that they are positioned similarly in relation to the stance object. The alignment may also be divergent, in which case they are positioned differently in relation to the stance object. The third point of the stance triangle is the Object. This is the dialogic entity which is being evaluated by the author, and which the author positions themselves in relation to. In summary, during a stancetaking act, the author evaluates a stance object, positions themselves in relation to that object, and expresses a degree of alignment with the intended audience, which may converge or diverge. Stancetaking relates to a number of pragmatic phenomena, such as humor, politeness, and expressions of irony (Riloff et al., 2013). As I suggested in Chapter 3, AOs often occur in contexts where irony or parody is a community norm, therefore irony is of particular interest to the study of AOs.

Irony has proven notoriously difficult to study, as it is difficult to identify the subtleties that characterize and produce irony. Irony has been classically defined as a type of meaning inversion, where what the speaker says is opposite from what they mean (Gibbs & Colston, 2007; Sperber & Wilson, 1981). Research suggests that people use irony to persuade, entertain, or build relationships (Colston, 1997; Dews et al., 1995; Pexman & Olineck, 2002; Pexman & Zvaigzne, 2004).

As irony and sarcasm are essential to understanding AO use, I sought to include irony as a key component of my approach to the study of stance. Irony has proven to be much more complex than simply producing opposite meanings, however, defining irony as a type of meaning inversion allows for the possibility of analyzing AOs along a spectrum from sincere to inverted, or ironic. Therefore, using the Du Bois (2007) stance triangle as a framework, I conceive of irony as a manifestation of positioning. I frame irony as a positioning act, in which the speaker distances themselves from the stance object. When irony occurs, it creates a "mismatch between the speakers' affective stance and the linguistic content of their utterance, revealing the intent of the speakers to *distance* themselves from the thought or idea expressed by the content." (Mauchand et al., 2020, p. 142). Rather than treating irony as something that is done on top of the stancetaking act, by framing positioning as a measure of irony, I make irony an essential part of enacting stance.

In the next section, I build on this idea of positioning as degrees of irony by proposing a typology of AOs along two axes of evaluation and positioning. Rather than suggesting that AOs either perform one function or another, I suggest that AOs are multifunctional and may serve multiple purposes at the same time. Instead, AOs exist on a spectrum of various meaning, with the degrees of evaluation and irony varying substantially across examples.

4.3 A typology of AO stance marking functions

In the last section, I established that contextualization cues have no conventional meaning, and context is critical in deriving their pragmatic contributions. Because of this, objectively characterizing the functions of AOs is difficult. One of the most challenging aspects of analyzing AOs is in effectively capturing the multifunctionality of these operators. In Williams (2021), I demonstrated that AOs construe evaluation, signal affiliation, and enable critique, but in many cases these operators enacting multiple functions simultaneously.

In this section, I propose two spectrums which effectively capture the wide ranging functions of AOs. The goal is not to attempt to place AOs into discrete categories, but rather explain the range of functions which they may enact. As I am framing AOs as stance markers, I co-opt the two speaker-oriented components of the Du Bois stance triangle, and place AOs along these edges. The first edge is evaluation, which connects the author and the stance object. I argue that AOs exist across a range of evaluation, with some AOs being very clearly evaluative and others seeming to lack any clear evaluative meaning. The second edge is positioning, which also connects the author and the stance object. This edge ranges from fully ironic or distancing, to fully sincere or matching. Again, I would like to emphasize that these are not discrete categories, but instead a continuum for examining the full range of possible functions that AOs may enact. By looking at AOs through the lens of these two dimensions, we may discern what features help distinguish them from one another (and likewise distinguish the functions they enact).

4.3.1 Evaluation

The first dimension I explore is evaluation. For the purposes of this typology, I define evaluation as occurring when the AO contains text that indicates an attitude, emotion or other subjective information. Following this definition, the affixed text that an AO modifies may range from + evaluative to – evaluative. For instance, examples 22-24 contain AOs that are + evaluative.

22. **Make Television Great Again™**
r/AskReddit

23. Signalling is for us **Poor™** people
r/AskReddit

24. One night, I had a particularly **Bad Feeling™** as I was trying to fall asleep…
r/AskReddit

I label these AOs as evaluative as they express the author's perspective and contain subjective information. With evaluative adjectives such as *Great*, there is not a clear definition for how *great* something must be to merit calling it *great*. In other words, if we take the lexical items as *x*, there is no clear threshold for how *x* something must be in order to be called *x*. The particular author's perspective is required to interpret the meaning. Therefore, I refer to these as + evaluative. By contrast, examples 25-27 are examples are what I term – evaluative.

25. I was raised Mormon. I was seminary class President, on the stake youth council, in leadership in my ward, basically the whole **deal** ™. Now I'm a swearing, drinking, weed smoking lesbian living in "sin" with my girlfriend
r/AskReddit

26. **Subway** ™ Sandwich is so good it gives me a heart attack and I die of shock
r/AskReddit

27. Those aren't your cup of Earl Grey **tea**™? Then swing by **Wendy's**™ and try something off the value menu! Great savings, great tastes, great fun!
r/AskReddit

These examples lack the attitudinal nature of examples 22-24. While these examples may require context to fully understand the author's intentions, the affixed units do not yield the same degree of author evaluation. In example 27, a cup of tea is either a cup of tea, or it is not. There is no perspective required. As such, I call these – evaluative.

Note that this does not mean that the utterances above do not construe any evaluation or take any part in stance marking, but whatever evaluation is occurring is not inherent in the affixed text itself. Furthermore, as I have placed these on a spectrum rather than placing them in discrete categories, there is room for ambiguity and perspective in terms of defining these terms (e.g., what counts as *tea*? what counts as a *deal*?). From a certain philosophical perspective, all language use is subjective. However, the main point I wish to make here is that there are certainly some affixed units which are more evaluative, and those which are less so. Following from this, there are many examples that do not fit cleanly into either category. Take example 28 below.

28. Ahhh, that **New Car Smell™**
r/AskReddit

In 29, upon first glance, there does not appear to be anything overly subjective about the affixed unit, *New Car Smell*. However, upon closer examination, there is room for subjectivity. How *new* does a car have to be in order to have the *new car smell?* How strong does the smell need to be in order to have the *New Car Smell*? Ambiguous cases like this are why I propose a continuum rather than discrete categories for evaluation.

4.3.2 Positioning

The second dimension I propose for describing AOs is positioning. As I previously described, I frame positioning as a way of conceiving of sarcasm or sincerity. If an affixed unit is – positioned (e.g., distanced), then the affixed unit meaning has been inverted or distorted from what it otherwise means. If an affixed unit is + positioned (e.g., matched) then there is no overt meaning inversion and the meaning is aligned with the propositional content of the affixed unit. As with the evaluation dimension, I propose this on a spectrum rather than discrete categories. This is partially because of the reasons stated previously, but also because meaning inversion is quite nuanced and may be interpreted different ways by different people in particular contexts. Some comments in which AOs are – positioned are below in examples 29-30.

29. You wouldn't think so, you are obviously a pro-abortion, homosexual communist. If you were a **Real American™**, you would have found it to be hilarious.
r/xkcd

30. He wasnt a **True Christian©** according to the Catholic Church and he got thrown to the fire... seems like the point still stands.
r/atheism

In example 29, the context suggests that the author does not actually use the phrase *Real American* the way that they themselves would define *Real American*. (E.g., pro-abortion, homosexual communist). Instead, they are borrowing and even parodying a definition proposed by someone else. This is even more clear in example 30, when the author attributes the definition of *True Christian* to the Catholic Church. In affixed units that are – positioned, sarcasm is often in play, and the affixed unit is often imitating another party's description or definition. In situations where the affixed unit exists in the – positioned plane, the author is distancing themselves from the affixed unit or stance object. Additional instances of this are below in examples 31-36, with my proposal for the effective meaning underneath the text below.

31. That's just god trying to show you the **Truth®**
r/atheism

--which I do not really believe is the Truth

32. He wasn't a "**True Christian©**" according to the Catholic Church and he got thrown to the fire... seems like the point still stands.
r/atheism

--I do not believe True Christian is a valid designation

33. Those people are probably ignorant, hell they might be racist, but they aren't liars nor did they run on a platform of **"change"** ©
r/technology

--*I do not believe there was meaningful change*

34. Well, the war certainly settled the issue didn't it?

Violence©- Solving Your Problems Since Cain smacked Able**
r/history

--*I do not actually believe Violence solves problems*

35. So by procreating, straights are fostering the **Gay Agenda™**
r/gay

--I do not think the gay agenda is a real thing

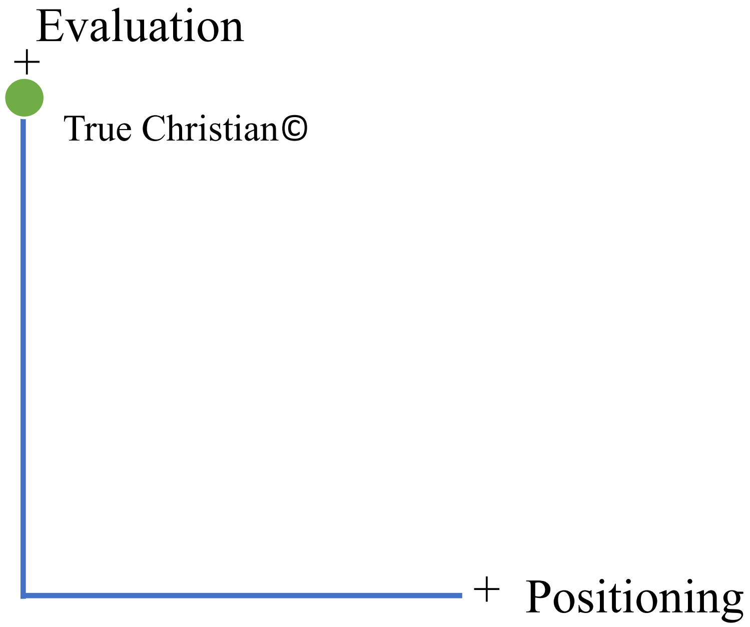
36. Being gay is not natural. **Real Americans ™** always reject unnatural things like eyeglasses, polyester, and air conditioning
r/atheism

--I do not believe there is such a thing as Real Americans

In general, in all the examples above, the author is distancing themselves from the content of the affixed unit. Therefore, I posit that one of the core stance-marking impacts that the artificial operator has is distancing the author from the utterance.

In summary, I have proposed a two-dimensional typology for describing affixed text modified by AOs. An affixed unit may be +/– evaluative and +/– positioned. The combination of these two dimensions yield often complex contextual meaning. *True Christian* from the previous examples has been placed on an example mapping of these two dimensions below in Figure 4.2.

Figure 4.2: Typology for capturing AO functions



True Christian is labeled as + Evaluative. This is because the phrase *true Christian* is subjective and may yield to disagreements over the criteria that merits being a genuine or ‘good’ Christian.

From the context of the full comment, I also argued that the author did not actually subscribe to the definition or even existence of a category *True Christian*. As such this is also – Positioned. Therefore, the author simultaneously produces evaluative content, and then proceeds to distance themselves from it, to ensure readers know that it is in not in line with their opinions. Instead, they take a stance that mocks the perspective they are imitating.

In this section, I have noted that the two speaker-oriented stance-marking components of evaluation and positioning are enacted simultaneously and to various degrees. Rather than proposing that these dimensions capture all the functions of AOs, this is merely one framework for capturing the multifunctionality of AOs. For the remainder of this chapter, I will explore these and other dimensions in more detail, relating them back more explicitly to theories of stance.

4.4 Evaluation

Returning to the stance triangle, a fundamental component of every stancetaking act is the evaluation of a stance object. In this section, I explore the evaluative nature of AOs in more depth. I look first at evaluative metacommentary, as a well-established function of the hashtag. I then look at ways that users on Reddit construe evaluation through overtly + evaluative AOs.

4.4.1 Evaluative metacommentary

Evaluative metacommentary is a well-established function of the hashtag (Zappavigna 2012, 2015, 2018; Wikström, 2014). Williams (2021) demonstrated that evaluative metacommentary is also a function of the trademark as in examples 37-38 below.

37. wow Sasha was top and trinity bombed???? **Gagged™**

38. We getting it, don't worry. **Soon™**

(adapted from Williams, 2021, p. 171)

The target of the evaluative metacommentary may be the post itself (intra-textual) or the context in which the post occurs (inter-textual). In studies of the hashtag, evaluative metacommentary is the most obvious way that they are utilized to express attitude and stance of the speaker. Evaluative metacommentary has been analyzed extensively as a function of the hashtag. However, in addition to evaluative metacommentary, the dataset revealed AOs being involved in construing evaluation more broadly. I discuss these in the next section.

4.4.2 Overt evaluation

Overt evaluation are those examples which would be labeled + evaluative along the proposed typology. They contain explicitly attitudinal content in the affixed text. I distinguish overt evaluation from evaluative metacommentary for a few reasons. The first is that evaluative metacommentary typically occurs at the end of the post. By contrast, examples of overt evaluation may occur anywhere within the sentence. Additionally, metacommentary specifically directs evaluation toward the post or content itself, whereas overt evaluation is less limited. These are situations where the role of the AO is to enhance or strengthen the meaning of the evaluative affixed unit.

39. It's manipulation. Added with a false ultimatum. She's only upset because you didn't do what she expected you to do. That's not the kind of shit you keep around yourself in a healthy **relationshit©**.
r/AmItheAsshole

40. We should be happy that being a minority isn't really a thing precluding someone from office anymore. That in itself is a bigger achievement than it might feel for us younger folk who have grown up not as exposed to all the crap that used to prevail with people who are now in late 40s and 50s. The voting majority has shifted with history, **A Good Thing®**
r/SandersForPresident

41. I am now the proud owner of the original artwork of this comic! My wife is **awesome™**.
r/comics

In each of these examples, the affixed content is overtly evaluative. In example 39, the author makes a play on the word *relationship* by changing the latter half of the word into an expletive. In examples 40 and 41, the examples are positively evaluating in that they contain the words *good* and *awesome*. We may interpret these examples using the stance triangle. In example 39, the stance object that is being evaluated is the unhealthy relationship, which the author terms a *relationshit*. This is also true in examples 40 and 41, with the voting majority shifting with history and the author's wife being evaluated, respectively. In these overtly evaluative examples, a stance object is evaluated, and the role of the operator appears to be enhancing that evaluation.

In addition to these overt examples, authors may construe evaluation using AOs without using overtly evaluative language. However, construing evaluation in this way requires contextual information and understanding of how the speakers are positioned in relation to the stance object and their audience. This kind of evaluation is more complex and may be related to shared knowledge among the subreddit community. In other words, the affective stancetaking happens via a complex system of context activation via contextualization cues, leveraging context to understand the author's positioning, and then interpreting their evaluation of the stance object. I will discuss this more general construal of evaluation via affiliation and criticism in the next section. I will also show the important role that alignment plays in enabling evaluation to be conveyed.

4.5 Positioning

In the previous section I discussed matching and distancing as ways that an author can position themselves in relation to a stance object. Another way of describing this is answering the question: to what degree does the author mean what they say? With AOs, this primarily becomes relevant as a critical distancing function, where the author marks a disconnect between their position and the position described within the sentence or affixed text. This is similar to what Cirillo (2019) calls ‘managing attributions.’

I demonstrated in the previous section that distancing is often a function of AOs. This distancing occurs when the AO signals a mismatch between the affective stance of the author and the propositional content of what they have written. In this section, I present collections of AO comments from different subreddits. As context is an important component of the meaning, I have chosen to present groups of examples of the same AO from the same subreddit. The first group of examples are from the banned subreddit r/The_Donald, which was a pro-Donald Trump subreddit.

42. Good news is that since dogs are absolutely haram this will ease your conversion to the **Religion of Peace©**

43. No sir. That was another cultural enrichment open class proudly brought to you by **The Religion of Peace©**. And now, a word from our sponsor, Open Society NGO.

44. Yep, Fake News is right folks. **The Religion of Peace ©** has no part in those attacks. It's those crazy Buddhists again....

45. Ehh. Unless I'm mistaken the local government seems to dislike Christians as well as Muslims. Them ignoring warnings is somewhat uncomfortable.

    I forgot to add: Oh Em Gee. **The Religion of Peace©?** Surely not!!11!!cos(0)!

In the examples above, members of r/The_Donald refer to Islam as *The Religion of Peace©*. In these cases, the copyright is performing a critical distancing function and signaling a mismatch between the propositional content of the affixed text and the affective stance of the author. Specifically, the role of the copyright symbol is to indicate that the author does not actually believe Islam to be a religion of peace, as evident by the references to attacks in example 44. In this way, the basic role of the copyright symbol is to critically distance the speaker from the propositional content of the text. This distancing in turn creates an evaluative effect. Throughout these examples, the speakers are negatively evaluating Islam, the stance object.

The registered trademark symbol is used in the corpus to achieve the same effect. This group of examples comes from the subreddit r/conspiracy. This is a community where people talk about conspiracy theories. In these examples, they reference the events of the Holocaust.

46. Remember, any questioning of any aspect of the **Holocaust®** is automatically hate speech.

47. Hitler's real crime was stopping Jewish subversion and holding them accountable. This is why the **Holocaus**t® gets hyped more than any other aspect of WW2. Particularly, when the exact same type of subversion is currently occurring in the US.

48. Without the **Holocaust**®, there would be no Israel. The official **Holocaust**® narrative is the basis used to suppress any opposition to their agenda and exposure of their criminality.

49. The article answers your question. Jews take it as an insult to question their Official Story of the **Holocaust**®

50. The **Holocaust®** has been exaggerated in order to portray the Nazis as the embodiment of evil. Why? To justify the need for Israel and to ensure that no other nation attempts to hold Jews accountable for subversion. So far, it has worked brilliantly.

These examples are produced by a group of people who appear to believe in conspiracies surrounding the holocaust. As such, they distance themselves from the history of the holocaust by affixing the word with the registered trademark symbol. Similar to the last group of examples, we see critical distancing happening in large part due to the presence of the AO. The registered trademark symbol signals to the readers that when they refer to the holocaust, they do not agree with the meaning invoked by it. In this case, the affective stance of the speaker is negative and the stance object is the *exaggerated* version of the holocaust. Here, the distancing effect allows them to refer to a series of events that they don’t believe happened, simultaneously ensuring that readers know what they are talking about, and simultaneously signaling that they are skeptical of the widely accepted historical facts.

The last group of examples of critical distancing I will present in this section are also from the subreddit r/The_Donald. This group features the trademark symbol being used to the same effect, with the phrase *Tolerant Left*.

> 51. So the **Tolerant Left™** burned down a church? Nothing these people do surprise me anymore. Anything for that vote fuck everything else.
>
> 52. I'm sure the **Tolerant Left™** will condemn tonight's attempt on Donald's life (they won't).
>
> 53. Careful, don't forget any dissenting opinions are now hate speech.
>
> Brought to you by **The Tolerant Left™**
>
> 54. Remember when the **Tolerant Left™** told Lil Wayne that he didn't know what he experienced as a black man and that they (white cuck liberals) did? Good times.
>
> 55. Well, considering 100% of Clinton supporters on Twitter are blaming white men for the results, I'll say that being less racist than the **Tolerant Left™** is a good start.
>
> 56. Hahaha. **The Tolerant Left™** rears its ugly head. Couldn't make a point so you resort to insulting my religion. Go ahead. Doesn't bother me one bit. Just proves the hypocrisy of the Clinton campaign and her supporters.
>
> 57. Remember when the girl who complained about Trump causing bullying in her school was found out to be a child actor? Just another case of projection by the **Tolerant Left™**.

Once again, in these examples there is meaning inversion and critical distancing from the propositional content inside the affixed text. Here it is rather clear that the writers are referring to the *Tolerant Left* with a negative evaluation toward the left and actually indicating their belief that the left is not tolerant. They leverage the trademark symbol to enact sarcasm and position themselves in such a way that it is clear to the reader that there is a mismatch between the words that they say and the affective stance they wish to convey.

In cases of critical distancing, the AOs play an important role in correctly conveying the attitude and affective stance of the author. However, rather than explicit evaluation, as we saw with the previous examples, the AO enables more nuanced communication via meaning inversion and signaling the mismatch between the propositional content and the stance of the author.

4.6 Alignment

Thus far in this chapter, I have not touched very much upon the last edge of the stance triangle: alignment. This edge conveys the degree of alignment between the author and their intended audience. This alignment may converge, if the two are aligned, or diverge if they are not. AOs as stance markers are capable of indicating converging alignment, or affiliation, as well as diverging alignment, or criticism.

4.6.1 Affiliation

Zappavigna (2012, 2018) talks extensively about ‘ambient affiliation’ and demonstrates that one function of the hashtag is creating this affiliation in online spaces. This affiliation occurs when hashtags are employed to indicate converging alignment with their intended audience. While affiliation is of course possible without the presence of a hashtag, Zappavigna effectively argues that the hashtag acts to enhance affiliation, and that users employ hashtags as one method to accept or reject bonds. Even standing alone, hashtags may signal a stance with which an audience may converge or diverge. As this relationship in a social media environment is often with an unknown or hypothetical audience, Zappavigna terms this ‘ambient affiliation.’ The basis for ‘ambient affiliation’ comes from dialogic affiliation (Knight, 2010). Dialogic affiliation is a framework for studying the bonds between conversational participants. Knight proposes that throughout conversations, interlocutors propose bonds, and use three strategies for affiliation in response to these proposed bonds. These response types are communing (sharing the bond), laughing (diffusing an unshared bond) and condemning (rejecting an unshared bond). Zappavigna adapts this model to a social media context, where the audience is dynamic and indeterminate. However, she demonstrates that hashtags, alongside other strategies, may help to accept or reject proposed bonds.

As with many other functions of the hashtag, I argue that signaling or creating affiliation is a function of other AOs, as well. Reddit is a particularly interesting dataset for examining affiliation, as content is organized into subreddits with a particular theme or topic. This creates some degree of affiliation inherent to the subreddits themselves. However, affiliation is more than simply marking community membership within the same subreddit. In the context of AOs, I argue that affiliation takes two forms. The first form is familiarity marking.

4.6.1.1 Familiarity marking

Familiarity-marking occurs when an author signals to the audience that there is shared, backgrounded information contained within the affixed text. In other words, in cases of familiarity marking, the AO signals to the reader that they are likely familiar with the propositional content within the affixed unit. I propose this as a type of affiliation because it assumes community membership and shared background knowledge of what the hypothetical audience should be familiar with as a member of that community.

The first group of examples for familiarity marking come from the subreddit r/niceguys.

58. Bro, this sub is for **NiceGuys©**, not for actual good hearted people.

59. Hey hes not that hard to understand, probs english is his 2nd language... but defs a **#niceguy**

60. Poor Jake. Poor dear, entitled **NiceGuy©** Jake. It's so unfortunate that these horrible ho-bags didn't see the magnificence of his outdated 90s emo sadness/needs/fauxdepression.

61. You can pretty much pick a **NiceGuy©** as soon as they use the word 'females' when talking about women.

62. Always better to be a nice guy than a **NiceGuy©**

63. 6 Harsh Truths That Will Make You A Better Person, an article I think every **#niceguy** should read.

64. Her (male) friend said "Let me chime in to summarise why **Nice Guys©** are inherent misogynists -- they believe women are obligated to date them because they're such 'nice' people and they can't understand that women might turn them down [and simply want to be friends] because they're just not interested. Nice guys, on the other hand, are nice to women because they're nice and respectful to everyone, without expectation that women should date them just because they're nice."

65. Is not taking a hint all it takes to be classified as a **#niceguy** though?

Familiarity Marking as affiliation works on two levels in these examples. The first is that there is an assumption that the hypothetical audience shares the definition and understanding of what is meant by *Nice Guy©* and *#niceguy*. Rather than meaning a guy who is nice, *Nice Guy©* and *#niceguy* refer to a collection of features including being *inherent misogynists* who *believe women are obligated to date them.* In example 62, *Nice Guy©* is directly contrasted with *nice guy*, demonstrating a clear difference between the affixed meaning and the unaffixed meaning. However, in addition to understanding the meaning of *Nice Guy©* and *#niceguy*, affiliation is occurring through the shared negative evaluation of the stance object: *Nice Guy©*. In this way, the AO does the work of communing affiliation through signaling both a shared of understanding of specialized language, and a converging alignment about the stance object between the author and their imagined audience.

In the politics subreddit they build up affiliation through familiarity marking by talking about tweets by Donald Trump with the registered trademark symbol.

66. Literally a **trump tweet®** for every occasion: [Only the Obama WH can get away with attacking Bob Woodward.] (https://twitter.com/realDonaldTrump/status/307582196196188160)

67. Keep this **trump tweet®** handy for when trump needs to appoint his third Chief of "Staffs" and third national security advisor in less than 1.5 years. https://twitter.com/realDonaldTrump/status/156829591267328000

68. He's trying to get him to quit so he can recess appoint a sycophant that will fire Mueller.
edit: because there is literally a hypocritical **Trump tweet®** for every occasion:
* ["I'm loyal to people who've done good work for me." #TheArtofTheDeal] (https://twitter.com/realDonaldTrump/status/194769587613605889)

69. I can't wait for this **trump tweet®** to go viral when he eventually falls down some stairs;
[Obama should stop running down the stairs when getting off Air Force One. Doesn't look presidential and at some point he will take a fall.] (https://twitter.com/realDonaldTrump/status/448104820176859136)

70. Incoming hypocritical **Trump tweet®**:
[President Obama refuses to answer question about Iran terror funding. **I won't dodge questions as your President.**] (https://twitter.com/realDonaldTrump/status/761386080272875520) ~ 7:19 PM - 4 Aug 2016

In this subreddit, people share relevant tweets from Donald Trump's twitter in threads on various topics. As with *NiceGuy©*, *Trump tweet®* is used not simply used to refer to a tweet written by trump, but a tweet with certain additional contextual qualities such as being hypocritical or inaccurate. As before, familiarity marking functions as affiliation by indicating a shared understanding of what specific qualities are entailed by *Trump tweet®*, but also a shared negative evaluation (and therefore shared alignment) and positioning toward the stance object: Trump's tweets.

The last group of examples of familiarity marking come from the subreddit r/iamverysmart.

71. yup, that's me also. I still do **verysmart™** things, like I simply had to post the first stage of the falcon 9 landing and explaining why it's important. BUT I did it in a not smug way! baby steps.

72. Lol OP you're way more **verysmart™** than he is. Just needed to take a few minutes of looking at your profile and you're a pretentious twat

73. Nah dude you good. We can't all be **#verysmart**

74. I don't even think she was making fun of anyone or out to put other people down, it had no context whatsoever. It was like she sat down and thought "What **verysmart™** thing can I say on Facebook to let everyone know I'm **verysmarrt™**".

75. Not really sure how this is **verysmart™**. Maybe abit cynical, sure, but he's not really that far off-base.

76. Why? Just check out the subreddit, it's cool people solving riddles and being **#verysmart**, it's quite straightforward

77. I don't think this is **verysmart™** material. He is just discussing something instead of boasting about his IQ or something along those lines

78. I'm intellectual intelligence. **-verysmart™**

79. context, she posted a cool video of sting ray and made the error of calling it a "manta ray." Paragraphs of **#verysmart** correction by Red followed.

80. Funny how people who claim to be **#verysmart** always happen to make embarrassing grammatical errors

In this group of examples, members of the r/iamverysmart community use the phrase *verysmart™* and *#verysmart* to refer to a specific set of qualities that go far beyond the semantic meaning of the phrase *very smart*. Rather than being ‘very smart,’ *verysmart™* and *#verysmart* are associated with features such as *boasting about IQ*, being *pretentious* or *smug*, and letting *everyone know* that you are smart. Within the community, affiliation is built up via a shared negative evaluation of people who have *#verysmart* qualities and a shared understanding of the specific qualities to which *verysmart™* refers.

4.6.1.2 Prototypicality marking

The second version of affiliation that I would like to propose is prototypicality marking. This is a subset of familiarity marking, in that it still fundamentally assumes backgrounded familiarity with propositional content. However, in the case of prototypicality marking, rather than simply pointing to a common or backgrounded entity, the AO indicates reference to a specific, idealized version of the affixed unit. In order to appeal to a 'prototype' or 'stereotype' version of an entity, a certain shared understanding of the world is assumed. Therefore, prototypicality marking is a type of affiliation in that it indicates shared viewpoints and common knowledge.

The first group of examples of prototypicality marking come from the subreddit r/atheism.

81. What moment made you realize god just isnt possible or likely

    When every religion claimed to be **THE TRUTH©**.

82. I was never religious, but my earliest bafflement moment was learning that most religions proclaimed to be the one and only **TRUTH©** and all others false.

83. Which claim does you denomination reject? That it has a corner on **THE TRUTH©**, or that homosexuals are going to hell?

84. "I'll pray for you." - to see the light, of course. The light of **TRUTH©**, which only they possess

85. I was never religious, none of them ever convinced me with their silly stories, each one claiming to be **THE TRUTH©**. It always seemed like nonsense, even as a kid.

86. Every religion purports to have the corner on **THE TRUTH©**

As with all examples of familiarity marking to create affiliation, *THE TRUTH©* here is associated with specific qualities that distinguish it from the unaffixed *truth*. However, what distinguishes prototypicality marking from other types of familiarity marking is that whereas a phrase like *verysmart™* has qualities associated with it that have nothing to do with the semantic meaning of the words *very smart*, in prototypicality marking the affixed text is being distilled into a quintessential, idealized form. In these examples, then, *THE TRUTH©* refers to the quintessential, or ultimate truth. As with all affiliation-related AOs, there is a shared evaluation and positioning to the stance object, which in this case, also relies on critical distancing. Therefore, the AO is simultaneously enacting prototypicality marking (the ultimate, quintessential truth), critical distancing (I do not believe in this ultimate, quintessential truth) and affiliation (my audience aligns with me and also does not believe in this ultimate, quintessential truth).

The second group of examples of prototypicality marking come from the subreddit, r/politics. In this case, they are affixing the word *Freedom* with the registered trademark symbol.

87. A small amount to pay for killing a quarter of that in the name of **Freedom®**

88. You just don't understand all the **Freedom®** that we have.

89. Side effects of **Freedom®** may include unnecessary war, waterboarding, covert actions and blowback, violence in neighboring countries, and the world's highest incarceration rate. Ask your conscience if **Freedom®** is right for you.

90. Well you took a big bite of the propaganda sandwich didn't you? Does it taste like **Freedom®**? You are far more likely to be killed by a cop. Direct your paranoia appropriately.

91. They don't like their **Freedom®** flavored rivers?

92. Dude, how young are you? Do you not recall hearing how "they" hate us for our "**Freedom®**" like a zillion times in order to justify invading foreign countries?

Freedom® here refers to the quintessential, or perhaps stereotypical notion of freedom. From context, it appears to refer specifically to the prototype of *Freedom* associated with the United States of America. As with the last group of examples, critical distancing occurs, with the AO signaling that the authors do not believe in the *Freedom®* which they are referencing. So once again, the AO serves three purposes simultaneously. Prototypicality marking signals that the author is speaking about quintessential, stereotypical *freedom*. Critical distancing signals that the author does not actually believe in that type of *freedom*. Affiliation is built by assuming that the audience aligns with the author and also does not believe in that type of freedom.

In summary, authors use AOs to commune affiliation by marking entities as backgrounded and familiar. They alter the meaning of text to mean more than the semantic component parts by signaling either that they mean something else with a very specific set of qualities, or that they mean a quintessential prototype of the entity. Crucially, affiliation happens when authors leverage these AOs to assume alignment between themselves and their audience with regards to the stance object they are evaluating. However, AOs do not always work to signal affiliation. In the next section I will discuss instances where AOs are used to signal divergent alignments.

4.6.2 Criticism

When an author takes a stance, they may converge or diverge with their hypothetical audience in terms of alignment. In this section, I examine how AOs are used in stancetaking to enact criticism. Zappavigna (2018) demonstrates that within the realm of affiliation, hashtags may be used to achieve two subtypes of criticism: censure and ridicule. Returning to the idea of affiliation as rejection or acceptance of bonds, censure may be defined as an author rejecting a bond or explicitly critiquing a stance object. She points out that while these inherently reject bonds with the individuals who are targets of the criticism, they are also aligning activities for individuals whose stances converge. Sarcasm, as a form of critical distancing, is a very important component of criticism, particularly of Ridicule, which involves the author ironically imitating the stance which they are critiquing, mocking the stance. As with hashtags, the AOs I am studying on Reddit are leveraged to enact both censure and ridicule.

4.6.2.1 Censure

Examples of censure come from the subreddit r/politics and involve authors affixing the name Trump with a copyright symbol as in the examples below.

93. "It's the president's right to spend $24 million to upgrade Air Force One refrigerators so they can humbly stock **Trump©** brand steaks." ~also Sean Hannity... probably.

94. Never ending lies by **Trump©**

95. One official **Trump©** Brand "Goodboye" token.

    Get them while you can! They won't be available much longer...

96. Would've been so much more meaningful if he'd brought **Trump©** Paper Towels.

In the above examples we see authors affixing the copyright symbol to Trump’s name, to create critical accusations that Trump used the presidency for his own financial gain. Example 93 posits a hypothetical situation in which Donald Trump spends millions of public dollars so that he can use public money to buy his own steaks. Similarly, in example 96, the author mocks the idea that Trump will use any excuse to promote his own products. Examples 94 and 95 are less overtly about a specific Trump product, but still use the copyright symbol as part of a broader censuring act for Trump’s financial misdeeds.

The next group of examples comes from the subreddit r/Buttcoin, a community of Bitcoin skeptics. In these examples, authors use the registered trademark symbol to censure and criticize the concept of cryptocurrency.

97. Remember that time you went to transfer **Actual Money®** between your accounts, and the bank threw it into a bottomless hole instead?

98. I believe that the sobering realization you are describing only comes after much SFYL. Sadly, I suspect that this new bagholder will have to be punished many times before he surrenders to **Actual Money®**.

99. Probably only a couple billion of **Actual Money®** in the crypto space. Market cap is still vastly inflated and would completely collapse if everyone took their money out.

100. From one rube to another, I can agree with you - except for the part about not reading /r/Buttcoin, that notion is haram.

All I know is that bitsoin is a means by which fools constantly devise new ways to throw their **Actual Money®** into a giant furnace, while a small group of people quietly intercept this money before it lands in the flames.

Through these examples, authors negatively evaluate the stance object bitcoin, by contrasting it with *actual money.* The role of the registered trademark here is to positively evaluate traditional currency by assigning it the status of *actual* or *real* money, thus implicating that bitcoin is not. This is an act of affiliation in that it diverges from the stances of people who are pro-bitcoin, but converges with other skeptics in the community.

4.6.2.2 Ridicule

Of the two types of criticism, ridicule is more subtle. It imitates an opposing stance, often in ridiculous ways as a means of mocking the stance. Therefore, by definition, ridicule always co-occurs with critical distancing and irony. Examples of ridicule come from the subreddit r/CringeAnarchy and involve the registered trademark symbol.

101. Typical leftist mentality. It's all but guaranteed that she didn't apologize for her ignorance either.

And if she actually did it was while blaming the **white male patriarchy of pure evil ®**

102. I wonder how the leftist media will spin this to blame the NRA and the **white male patriarchy of pure evil ®**

It should provide lots of cringe.

Or they'll just completely ignore it.

103. No doubt poor because of the **white male patriarchy of pure evil ®**

104. The **white male patriarchy of pure evil ®** strikes again!

105. They'll explain how the **white male patriarchy of pure evil ®** forced her to attempted murder and suicide.

In each of the above examples, authors affix the registered trademark to the phrase *white male patriarchy of pure evil.* In these examples, authors imitate an opposing stance that they characterize as unfairly blaming problems on a fictional *white male patriarchy* that is *pure evil.* Under their characterization, people who hold this stance do not distinguish among white males, do not believe there is nuance in the morality of white males, and unfairly blame things like murder and suicide on white males. As previously stated, critical distancing is a requirement for enacting ridicule. As with examples of critical distancing I analyzed at in the last section, authors use the AO to signal that they do not actually believe there is such a thing as the *white male patriarchy of pure evil.* Additionally, as part of the act of creating ridicule, they exaggerate the belief of the opposing view, imitating that exaggeration in order to mock and critique the stance they disagree with.

4.7 Conclusion

In this chapter, I have demonstrated that AOs enact a number of stance marking functions. I have shown that it is helpful to think of AOs in terms of their metadiscursive functions, granting text special information status. Furthermore, I have suggested that AOs be framed as contextualization cues, lacking conventional meaning but indicating the necessity of context.

I have discussed approaches to understanding stance, leveraging the Du Bois (2007) stance triangle as a useful way for analyzing stancetaking. I have proposed that as stance markers, AOs enact multiple functions simultaneously, and that it is therefore not helpful to place them in discrete categories. Instead, I have shown that stance marking AOs exist on dimensions of meaning, ranging in degrees of evaluation and positioning. Finally, I have shown that intersections of evaluation, positioning, and alignment yield higher order functions of affiliation and critique.

I adopt stance marking as an expression of speaker attitude and meaning. While I do not aim to suggest that in these examples, the AOs are the only element which influences stance, I have demonstrated in this section that they make substantial contributions to indicating author attitude and positioning via evaluation, critical distancing, affiliation and criticism.

CHAPTER 5

ARTIFICIAL OPERATORS AND SCALAR MODIFICATION

5.1 Background

In the previous chapter, I argued that a primary pragmatic function of Artificial Operators (AOs) is stance marking. While stance marking as a framework captures a wide range of AO meaning, in this chapter I will demonstrate that AOs also appear to interact with scalar meaning. Crucially, I do not propose that these functions are mutually exclusive – to the contrary I think that even AOs with scalar interaction often participate in stance marking. However, I posit here that the contributions AOs make when interacting with adjectives are an important dimension, which in many cases enhances and supports an AO's stance marking properties.

In this chapter, I am interested in two distinct concepts. The first is scalar modification with gradable adjectives. As noted in the corpus chapter, AOs co-occur with every part of speech category. While noun phrases are the most common, there are a substantial portion of AOs which co-occur with adjective and adverb phrases. AOs and adjectives and adverbs have a particularly compelling relationship, which I aim to explore in this chapter. I will argue that in the case of gradable adjectives, the AO adds semantic meaning by upscaling the adjective, as in example 106 below.

106. Roll for the Galaxy would be my top choice since it gives you a strategic, replayable euro game with great production value that *actually* plays in less than an hour. Also, dice rolling is **Fun™**.
r/boardgames

I argue that this meaning is semantically similar to adverbials such as *more, really, very, etc.* As with these terms, the role of the AO is to take the base adjective and move it to a higher place on the scale.

The second concept I discuss similarly looks at gradable adjectives, however it looks at scalar inversion. This occurs when rather than upscaling the effect, the role of the AO is to downscale or invert the meaning of the word. While I have argued that meaning inversion can happen as a function of stance marking, in this chapter I will provide evidence of meaning inversion being conventionalized to the point of widespread understanding that a term means something distinct from the un-affixed form, as in the example below.

107. They are, already confirmed that publicly in a post and/or on here. I also know they confirmed the new store will be **soon™** and not **soon**
r/Eve

The relationship between gradable adjectives and AOs is additionally interesting because gradable adjectives require invocation of a contextually-determined standard of comparison. AOs rely heavily on context to interpret meaning, and their interaction with scalar meaning adds a layer of complexity.

5.1.1 Discourse-oriented adjectives

Discourse-oriented adjectives have warranted substantial study from both semantic and pragmatic perspectives. Discourse-oriented may be defined as a circumstance when the truth of a sentence depends on perspective. In other words, discourse-oriented adjectives may be defined as those where there is no objective truth, and the meaning of the adjective relies on context. Discourse-oriented adjectives have been referred to in the literature by a number of different names including gradable adjectives, discourse-oriented adjectives, and subjective adjectives. Kennedy (2007, p.2) discusses the truth-conditions associated with utterances containing discourse-oriented adjectives . He refers to the sentence below.

108. The coffee in Rome is expensive.

Kennedy notes that this sentence introduces vagueness because "what exactly it means to 'count as' expensive is unclear." The difficulty of defining a standard for what counts as expensive may be illustrated by The Sorites Paradox (adapted from Kennedy, 2007, p. 2).

109.
i. A $5 cup of coffee is expensive (for a cup of coffee).
ii. Any cup of coffee that costs 1 cent less than an expensive one is expensive (for a cup of coffee).
iii. Therefore, any free cup of coffee is expensive.

This final conclusion is intuitively false, however the paradox demonstrates the difficulty in defining a clear standard for something as relative as being *expensive.* This kind of subjectivity is often diagnosed in the literature through 'faultless disagreement.' Compare this with more 'objective' sentences such as examples 110 and 111 repurposed from Kaiser & Rudin (2020, p. 698) below.

110. Arnold: The shirt is cotton.
Barbara: No, it's not cotton.

111. Amy: That knife is plastic.
Bob: No, it's not plastic.

In each of these examples, one of the two interlocutors is objectively wrong. That is, the knife is either plastic or it is not. The shirt is either cotton or it is not. These may be contrasted by the following subjective examples (also repurposed from Kaiser & Rudin 2020, p. 698).

112. Arnold: That rollercoaster was fun.
Barbara: No, it was not fun.

113. Amy: That sandwich was tasty.
Bob: No, it was not tasty.

In these examples, there is no objective truth. Neither of the interlocutors is wrong. Amy found the sandwich to be tasty, while Bob did not, but neither of them can be said to be incorrect. This phenomenon is referred to as faultless disagreement.

For the purposes of this paper, an adjective may be considered discourse-oriented if it can occur in cases of faultless disagreement. The core element that drives faultless disagreement with respect to these subjective adjectives has to do with the fact that they are standard-sensitive. That is, there is a contextual standard of comparison that the interlocutors refer to in order to make the claim that the rollercoaster was *fun.* Different rollercoasters may be less fun or more fun, but when they are labeled as simply *fun*, the interlocutor is stating that the rollercoaster was sufficiently high on the *fun* scale as to be labeled fun. In this way, these standard-sensitive adjectives refer to a relative scale of degrees of fun and mark the rollercoaster as meeting the necessary threshold to be considered fun in a particular context.

Within the realm of subjective adjectives, we may further distinguish between Relative Gradable Adjectives (RGAs) and Predicates of Personal Taste (PPTs). RGAs are adjectives that are standard-adhering but are not able to facilitate faultless disagreement in the comparative form. See examples 114-115 below.

114. A: Henry is taller than Prateek.
B: No, Prateek is taller than Henry.

115. A: Hector is older than Kaijuan.
B: No, Kaijuan is older than Hector.

In the case of examples 114 and 115, one person is objectively wrong. However, note that in the base form, these same adjectives may be used in a subjective way as they are standard-adhering. See examples 116 and 117.

116. A: Shen is tall.
B: Shen? Not really.

117. A: Wow, Marty is so old.
B: Oh please, no he is not.

In these cases, because there is some subjectivity about how *old* one has to be to be considered *old*, these adjectives are discourse-oriented.

In summary, RGAs may facilitate faultless disagreement in the base form, but not in the comparative form. RGAs may be contrasted with PPTs, which may facilitate faultless disagreement in both the base and comparative form. See examples 118 and 119.

118. A: Going to the theatre is more fun than going hiking.
B: No way, hiking is so much more fun!

119. A: The pepperoni pizza is tastier than the veggie one.
B: No, the veggie pizza is tastier.

In these examples, even with the comparative form there is no clear objective truth. Neither A nor B is factually incorrect.

For the purposes of analyzing AOs, these concepts are helpful in that they refer to a subjective scale in which interlocutors are negotiating the standard of comparison as well as a necessary threshold. The adjective *warm* invokes a scale from least warm to most warm, and when standard-sensitivity and subjectivity are involved, that scale is affected by contextual information. For example, one might say “It’s warm today” in the middle of winter. In winter, the standard of comparison is likely less warm than the expected standard if one was to say “It’s warm today” in the middle of summer. In addition to this standard of comparison, interlocutors negotiate the threshold. This is where faultless disagreement plays a role. Even if there is a shared standard of comparison, personal preference might still lead interlocutors to disagree on whether it is warm for winter or warm for summer. In the next section, I will demonstrate the interaction between AOs and gradable adjectives. I will argue that the AO contributes semantic meaning in the form of upscaling the gradable adjective on the scale.

5.2 Upscaling

In this section, I look specifically at situations in which the artificial operator affixes to a subjective adjective or adverb. I demonstrate that the artificial operator interacts with the standard to which these subjective adjectives and adverbs are implicitly held. This may be observed in example 120 below.

120. my friends had a party and didn’t invite me so I got drunk with my family instead. i then had takeaway I didn’t rly enjoy and watched sad movies the rest of the day. overall, **sad™** day.
r/AskReddit

In this section, I will demonstrate that the role of the trademark here is to indicate that not only does this instance of *sad* meet the standard, but it exceeds it. I will provide evidence of this function by walking through examples of a pair of gradable adjectives which are commonly affixed in the corpus: *good* and *bad*.

Good and *bad* are both relative gradable adjectives, which additionally meet the criteria of predicates of personal taste. That is, they facilitate faultless disagreement in both the positive and comparative forms. In the following examples, authors use the AO to systematically upscale the meaning of the affixed word *good*, to indicate a higher location on the scale. That is, *good*™ is higher on the ‘good’ scale than *good*. We see this in the examples below.

121. From experience, throwing yourself into the deep end like this is a really, really good way to learn the language. It will take a while, but like learning any skill, that’s the way it goes. You’ve got to suck at it before you’ll get **good™**
r/iOSProgramming

122. Must win. Gonna be real **good™**
r/canucks

123. Yes. Since they started being **good™** basically.
r/GlobalOffensive

124. I have no issue with them exercising judgment. But that is not what they did. They consulted the Bush administration first. They asked daddy first. *That* is

the issue.

In my opinion it is the hallmark of a **Good™** journalist to piss off the *most powerful* people occasionally. It means they're doing their job.
r/IAmA

125. I've always found playing angels to be a better source of dark humour. Just because they're **Good™** doesn't mean they have to be nice.
r/rpg

126. **Good™** and **Evil™**
r/AskReddit

127. The Argument from Evil, which is that God is by definition **Good™** and therefore anything he does is **Good™**. Allowing evil to flourish is **Good™**. Allowing babies to starve in Africa while helping millionaire athletes win their sports competitions is **Good™**. Blessing mega-wealthy televangelists while hardworking people sell their last possession to send their kid to school is **Good™**. It's all **Good™** because He is **Good™** and therefore nothing he does can not be **Good™** by definition
r/atheism

128. I mean they can't honestly expect the state to get into a debate about theology/morality and only allow monuments of **good™** mythological creatures or symbols.
r/atheism

129. Fiat currency is **Good™**, but only when inflation and deflation are kept in dynamic equilibrium with the value they are intended to represent via faucets and sinks, and the system works for all of its participants equally
r/economics

In these examples, the adjective *good* is enhanced by the presence of the AO. This is particularly evident in example 127, in which it refers to the pure, unimpeachable goodness of God. Similar to the prototypicality marking discussed in the stance marking chapter, when the adjective good is upscaled, it refers to an almost maximized version of the adjective. That is, *good*™ is at the highest point on the good scale.

We see a similar effect with the relative gradable adjective, *bad*. As with *good,* the role of the artificial operator is related to moving *bad* higher on the *bad* scale. In other words, on the scale of meaning, *bad*™ is worse than *bad*. In example 130 the author is discussing cronyism and nepotism. The author claims that they do not see it as Bad™ but could see it being *bad* in a political sense. In this way, Bad™ appears to be stronger than un-affixed *bad,* which is bad with caveats.

130. I'm not objecting to the idea that some people are in the positions they're in due to cronyism/nepotism. My objection comes from the notion that I don't know where I feel that that's a bad thing.

A father, posing the reigns of the family business onto his sons - is that nepotism? Yes, it's pretty much the textbook definition of it - but I have a hard time seeing that and identifying it as **Bad™**.

I definitely DO see the argument for it being bad in a political sense, or in the sense of a major, publicly-traded company, but how do you stop that and should you, in all cases? What if the family member has legitimately been participating in politics/the company and knows the ins and outs of the job and is by all metrics a legitimate candidate for the job? Put simply, I don't see a way to solve this problem without disproportionately and deliberately rewarding people who lack qualifications and experience, and I don't think that just giving people free money will put them right on par with people who have that much money. I don't think it's that simple, and I don't think it's the slam-dunk, silver bullet **Society Fixer™** that liberals champion it as.

Are there incompetent people in high positions? Yes, but I'd argue that's the exception - not the norm. It's also subjective - "incompetent" to you is almost certainly not the same as "incompetent" to me, especially when discussing the more cerebral and abstract tasks that high positions are expected to perform versus the set and specific tasks working class folks perform.
r/PoliticalDiscussion

There is a similar occurrence in example 131, where the author distinguishes between a *bad dog,* and a *Bad™ dog*.

131. i imagine that you were an angry doberman in a former life.

but a bad one, not like **Bad™**, but bad as in with mange and no teeth
r/bestof

132. Oh boy, we're censoring this book again?

Time to get out my book burning suspenders for the new age of censorship, because "reality makes me feel **bad™**".
r/news

133. OMG how awful! :(How do people not understand that overfeeding fish is **Bad™**?
r/Aquariums

134. Bear in mind you should lock it to ambient temp, cooler components will cause condensation to form and that is **Bad™**
r/watercooling

135. I know it isn't much but I've passed the tweet to two low level Google people that I know, and explained why it's **Bad™**. Gotta get people talking about this internally.
r/KotakuInAction

While in all of these cases, there are stancetaking actions going on which undoubtedly allow for more complex meaning, I would like to assert that at a base level, the AO adds an upscaling meaning. Setting aside the contextual information, such as author information, or community beliefs, the default or conventional meaning of the AO when affixed to an adjective is to upscale it. In the next section I will demonstrate a different case, where the AO paired with a particular affixed unit has led to a conventionalized meaning which actually inverts the scale.

5.3 Scalar inversion: soon™

Soon is a popular affixed unit in the dataset. In the corpus of AOs from Reddit, *soon* occurs with an artificial operator 36,068 distinct times across 34,848 distinct comments. *Soon* is also one of the longer-running popular affixes, first occurring with an artificial operator in the corpus in January, 2009 in the subreddit *r/technology*. *Soon™* became popular in gaming communities through the massive multi-player online role-playing game: World of Warcraft. An example of *soon™* from the corpus is below.

136. In other words, **[Soon™]**(http://www.wowwiki.com/Soon).
r/gaming

In this example, the user supplies a link with an explanation that refers to a wiki page for World of Warcraft. On that page, there is a definition of *Soon™*. The definition has been excerpted below.

> **"Soon™**: Copyright pending 2004-2021 Blizzard Entertainment, Inc. All rights reserved. "Soon™" does not imply any particular date, time, decade, century, or millennia in the past, present, and certainly not the future. "Soon" shall make no contract or warranty between Blizzard Entertainment and the end user. "Soon" will arrive some day, Blizzard does guarantee that "soon" will be here before the end of time. Maybe. Do not make plans based on "soon" as Blizzard will not be liable for any misuse, use, or even casual glancing at "soon."" (Soon™, 2022)

While the above excerpt is not an official definition, it provides valuable context for the use of *soon* with an artificial operator. The author of example 136 deliberately provides this context for other readers that explicitly outlines the ways in which the inherent subjectivity of *soon* is being leveraged and subsequently mocked. The definition specifically emphasizes the vagueness of *soon™*, noting that *"Soon™ does not imply any particular date, time, decade, century or millennia…"* While this definition is clearly tongue-in-cheek, it rightfully points out that *soon* is standard-sensitive and does not indicate a specific time. Like the aforementioned example to do with coffee being expensive, the determination of how *soon* something should be to be labeled as *soon* cannot be objectively determined. However, in addition to *soon* being indeterminate, the definition seems to emphasize that due to the inherent vagueness, *soon™* often violates the expected standard and therefore takes on an opposite meaning: *not soon*. If the Sorites Paradox yields a conclusion that free coffee is expensive, in this situation it similarly yields a conclusion that something that happens at *the end of time* may be labeled soon.

Following from this, the examples below are instances of *soon™* used specifically to refer to situations like those described in the wiki definition. That is, *soon™* used to mean *not soon*.

137. I'm sure it'll be fixed **Soon™**
r/gaming

138. Has the new client been updated to handle URF mode, or is that one of the "**Soon™**" things?
r/leagueoflegends

139. Don't worry, they'll fix it all real **SOON™**
r/windowsphone

140. I'm on PS4, could be anywhere from soon to **Soon™**
r/gaming

141. They said they are developing a new game. **Soon™**.
r/leagueoflegends

142. But in this case it isn't **Soon™**, it actually says 2016!! :O
r/leagueoflegends

Of particular interest are examples 140 and 142. Notably, these examples include definitions that distinguish *soon™* from *soon*. In example 140, the author writes that an event could happen anywhere from *soon* to *Soon™*. This author directly contrasts the meaning of un-affixed *soon* with affixed *soon™*. With the above context in mind, I interpret the differences as follows: *soon* without the AOs is the word on a normal scale that adheres to a reasonable, context-derived standard, whereas *soon*™ refers to the version which violates any version of an acceptable standard and ultimately means *not soon*. Similarly, in example 142, the provision of a date for an event (*it actually says 2016*) precludes the use of *soon™*, which as suggested by the author, requires vagueness and indeterminacy. Thus, *soon™* has two layers of meaning. The first is to signal indeterminacy, and the second is to signal the violation or adjustment of the implicit standard of comparison, with *soon* being sooner on the scale than *soon™*.

However, there are also examples within the corpus where the inversion of the previous examples appears to be absent. In these cases, the effect is to highlight the standard sensitivity without necessarily having the additional meaning of *not soon*. See examples 143 and 144 below.

143. I do, indeed. I'll post a review of my feelings on the car **soon™**
r/mazda3

144. Delta snorlax was added n the last patch. Not available in classic mode but i hope **soon™**
r/PokemonInsurgence

In example 144, the user says that they hope *Delta snorlax* will be available *soon™*. In contrast with the previous examples, in which the implication is that *soon™* means *not soon*, the genuine hope of the author to access the feature suggests that the user is not employing meaning inversion, but is instead denoting the standard sensitivity and vagueness of the term *soon* to emphasize the indeterminacy of the time in the future when *Delta snorlax* will be available in classic mode. Therefore, while I observe that *soon™* is often used to create sarcasm and meaning inversion, there are clearly instances where the artificial operator is used for the previously discussed function of emphasizing standard sensitivity.

Finally, it is interesting to note that in some small numbers, *soon* has been paired with the other AOs, as in examples 145-154.

145. Not available yet. Should be on **soon©**
r/blackops3

146. Can I get a collective "Awwww yissssss" I get off work **SOON©**
r/pcgaming

147. Lol at the black or white posts. Bottom line rito does some great things and on the other hand they really need to step up their game on other aspects. They listen and respond immediately to small things like the sion ult change, but actually important things that cost more money and time, like replays, we get the usual "**soon©**"
r/leagueoflegends

148. I can confirm that the single will drop on a day that ends with Y. **#soon**
r/Muse

149. I haven't found any details on date/time. Just their twitter post with **#soon**.
r/CaravanPalace

150. same, i tweeted him a while ago asking when is the wax dropping to which he replied **#soon**
r/deathgrips

151. I am remembering that we have been hearing "**Soon®**." since the launch of WP 7 five years ago. It seems further away than ever. At what point do you decide the ship has sailed?
r/windowsphone

152. Release Date: **Soon®**
r/starcitizen

153. **Soon®**, still working on console version
r/EliteDangerous

154. "**Soon ©**"

Are you kidding me?, it's Soon^^TM. that's it, i'm done. YOU HAD ONE JOB!
r/Warthunder

The vast majority (97%) of the ~20K *soon* affixed units in the corpus are affixed with the ™, but in the above examples we see authors using soon with the other operators to similar effects. Examples 147, 148, 149, 151, and 152 appear to invoke *soon* with meaning inversion (e.g., *not soon*). However, the rest of the examples are ambiguous or appear to function the way that we saw with *good* and *bad*: to highlight the standard sensitivity. As I have shown in the previous chapters, this is compelling evidence that these operators are used to similar effects. However, crucially, in example 154, one author corrects another, noting that it should not be *soon©*, but *soon^^TM*. This suggests that even if they are used to similar effects, some users do consider them to be different, or at the very least note that one is more conventional than another.

It should be noted that *soon* is not evenly distributed across communities. Of the 2,984 subreddits that contained *soon* as an affixed unit, 55% only contained one instance of a *soon* AO. The top 20 subreddits by volume of the affixed unit *soon* are below in Table 5.1. Each of these subreddits was manually examined, and all of them are related to either gaming or technology.

Table 5.1: Top subreddits by volume of soon affixed units

Subreddit	**Soon AOs**
leagueoflegends	1654
starcitizen	483
heroesofthestorm	363
Warframe	342
Eve	337
DotA2	300
EscapefromTarkov	289
windowsphone	282
Warthunder	270
wow	269
BattlefieldV	266
hearthstone	251
Guildwars2	238
StarWarsBattlefront	223
Planetside	217
pcmasterrace	207
Overwatch	190
FFBraveExvius	188
EliteDangerous	168
MechanicalKeyboards	167

In previous chapters, I established that subreddits often constitute Communities of Practice and proposed that AOs represent a particularly useful example of shared repertoire within communities. I wish to use this case study of *soon™* to reiterate this point. Table 5.1 shows that *soon™* is prevalent in communities dealing with gaming and technology. As I have shown in this chapter, *soon™* is often used in such a way that relies on contextual background information to create meaning inversion. The prevalence of *soon*™ in gaming communities, alongside that common meaning inversion, provides a clear example of shared repertoire within a Community of Practice. These kinds of AOs present a particularly strong example of this shared repertoire, because of the complex layers of meaning that are invoked when AOs are used, and the required community-based background knowledge to access those layers of meaning.

In summary, *soon*™ is one example of a subjective adverb where the affix appears to be interacting with the inherent vagueness of the meaning of *soon*. This has been merely one example of a standard-sensitive affixed unit and the way that authors employ AOs to highlight and in some cases subvert the implicit standard of comparison invoked by subjective gradable terms.

5.4 Conclusion

In this chapter I have proposed that AOs interact with discourse-oriented adjectives (and adverbs) to impact where the modifier sits on the contextual scale. I have suggested that in novel or less common AOs, the default effect is for the AO to upscale the adjective. I demonstrated this by looking at excerpts which included two adjectives: *good* and *bad*. I showed explicit contrasts that authors drew between un-affixed adjectives and their affixed counterparts. I have also shown that in some instances, the affix is paired with a text in order to perform scalar inversion. I performed a case study on the affixed unit *soon* to show that in most cases, authors use *soon* to mean: *not soon*.

I do not suggest here that the previously described stance marking analysis is not also at play here. There are certainly a range of stancetaking actions going on in conjunction with both upscaling and scalar inversion. However, I would like to propose that when AOs affix to discourse-oriented adjectives, in particular, authors leverage the inherent contextual uncertainty to create a heightened version of the un-affixed meaning.

CHAPTER 6

SURVEY RESULTS

6.1 Background

Artificial operators (AOs) are an understudied class of CMC cues. Other than the hashtag, the operators have attracted very little scholarly attention. In the previous two chapters, I have posited functions of AOs related to stance marking and scalar modification. However, due to the limited work done on these operators, and the inherently subjective nature of qualitative analysis, I sought input and ratings from recruited participants to corroborate and enhance the qualitative findings. The benefits of getting input from recruited participants are multiple. First, participant input lessens the likelihood that I, the researcher, will overlook or ignore an important function of these operators. Second, participant input provides a way to corroborate or bolster the claims I have made in the qualitative analysis about the various functions that AOs may serve. Third, participant input allows us to do statistical comparisons between functions and operators in terms of how they are perceived, and under what circumstances.

The task design follows previous studies in experimental pragmatics in taking an interest in the ways in which participants process the intended (as opposed to literal) meaning of an utterance (e.g., Bott & Noveck, 2004; Grodner et al., 2010; Kronmüller et al., 2014). In this case, the surveys sought to reveal the inferences associated with a particular pragmatic phenomenon: AOs.

Two studies were conducted with the hashtag, copyright symbol, trademark symbol, and registered trademark symbol. The first study was a free-response style survey with 40 items. Participants were asked to describe, in their own words, why the author had used the artificial operator. The study was designed to test whether participants would register a difference between ‘figurative’ and ‘genuine’ instances of AOs. The second study asked participants to rate the accuracy of various explanations for AOs using a scale from 1 to 5. The second study was designed to test whether the functions of the operators would be perceived differently from one another in controlled contexts. In the remainder of this chapter, I describe the data collection and results of these two studies. I conclude with a discussion placing the results in context with the other findings of this dissertation.

6.2 Survey 1: genuine vs. figurative

6.2.1 Design and participants

Survey 1 sought to test the hypothesis that genuine uses of AOs and figurative uses of AOs would be perceived differently by English social media users. This question is important to my overall analysis of AOs because I have posited that users are repurposing them for new, communicative uses in CMC contexts. Crucial to my suggestion that AOs should be studied together is the notion that they are collectively being used in new, overlapping ways. In my qualitative analysis, I posit functions that are separate from their ‘original’ purpose.

The survey was administered through Question Pro. Participants were presented with excerpts from Reddit and asked to describe in their own words why they believe the author used the symbol. The study had 40 items, each was an excerpt from Reddit including a single AO, with 10 items per operator. For each AO, 5 of the items were control items and 5 were experimental. Control items were those in which the operator was used in a genuine way. Experimental items were those in which the operator was used in a figurative way. Control and experimental items were selected by the researcher from the Pushshift Reddit dataset, using context to determine if the usage was genuine or figurative. Participants responded via free response text to describe why the author used the operator.

Two sample items are below. The full list of items may be found in Appendix C. Figure 6.1 shows one of the experimental items, while Figure 6.2 shows one of the control items. Items were presented in random order.

Figure 6.1: Example experimental item for survey 1

Experimental Condition: Artificial Operator used in a pragmatic way

> *The below text is an excerpt from a comment on Reddit.com. Please read the text and then answer the question below.*
>
> **As much of Africa is on the rise, the situations in Burundi and Central African Republic remain very #Bad**
>
> *In your own words, please explain why you believe the author used the highlighted symbol here.*

Figure 6.2: Example control item for Survey 1

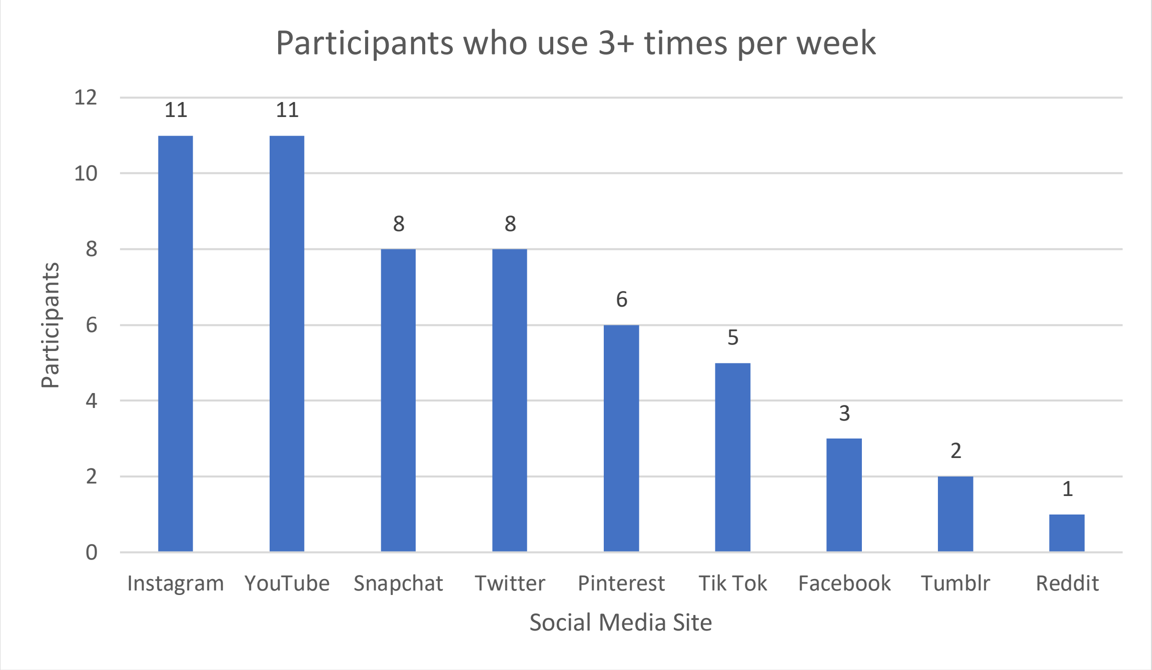
Control Condition: Artificial Operator used in traditional way

> *The below text is an excerpt from a comment on Reddit.com. Please read the text and then answer the question below.*
>
> **Antonio Palladino postet a photo of a tattooed guy one month ago with the hashtags #capetown and #tatoo. I don't think this is a coincidence.**
>
> *In your own words, please explain why you believe the author used the highlighted symbol here.*

15 participants were recruited to participate in the study. Undergraduate students enrolled in Linguistics courses at the University of Texas at Arlington were offered extra credit to participate. 11 women, 2 men, and 2 nonbinary people participated. All 15 participants were L1 users of English. 13 participants reported being between the ages of 18-24, 1 between 25-29, and

1 between 30-39. Participants were asked to report which social media platforms they use 3+ times per week. Participants could select multiple platforms. The distribution of people per social media site is reported in the figure below.

Figure 6.3: Survey 1 participant social media use



6.2.2 Results

Each of the 15 participants answered 40 free-response questions, yielding a total of 600 data points. Responses were hand-coded by the researcher using a bottom-up approach. That is, I created categories or codes in accordance with the data that I saw, and only created new codes if data points did not fit into the existing codes. After my first pass of coding, I noted those codes which only had one response associated with it or where the response was unclear. These were collapsed into a category: *Other*. If a participant simply wrote that they did not know, responses were coded as *No response*. Participants who attributed the function to be the genuine meaning of the AO were coded as *Literal*. The final list of codes is below in Table 6.1.

Table 6.1: Response coding scheme

Code	**Description**
Sarcasm	mark user intent of irony, mockery or critique
Quintessential	denote a prototypical instance of phrase
Emphasis	strengthen or draw attention to a phrase
Familiarity	mark a common, shared, known, instance of the phrase
Quotation	used to the same effect as quotation marks
Humor	mark humorous intent or language play
Literal	the genuine or default meaning of the AO
Other	unclear or does not merit its own category
No response	participant did not provide a meaningful response

It was possible for responses to receive more than one code such as the example below.

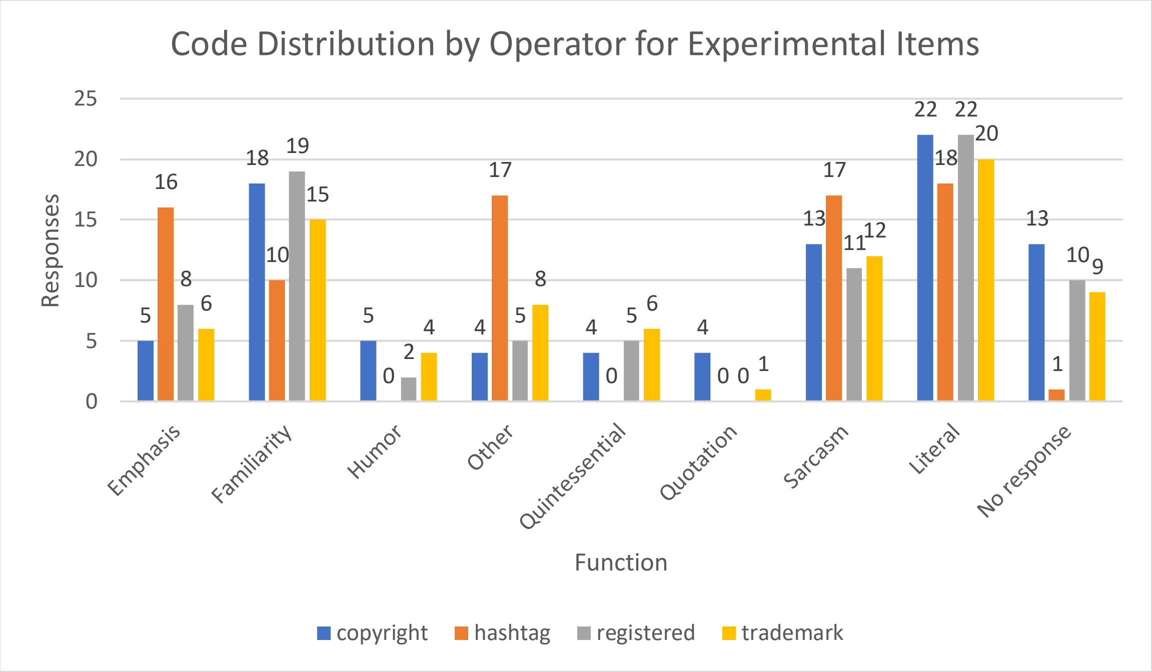
> 155. I think the author used the symbol to place emphasis on the idea of there being a 'true christian'. The use of the copyright symbol here feels sarcastic or ironic; we are (generally speaking) aware that there is this 'ideal version' of being a christian. It's use makes me feel like the author is mocking the idea of a 'true christian',and is adding extra emphasis to the quotation marks that are used here.

In this example, the response received three codes: emphasis, sarcasm and quotation. The participant states that the symbol places *emphasis* on the idea of *true Christian.* They specifically mention sarcasm and irony, Finally, they suggest that the symbol is interacting with or emphasizing the role of the quotation marks. In this way, many of the responses received multiple codes. In the remainder of this section, I will talk about the results for the experimental items and the control items. Finally, I share the results of the statistical analysis comparing the two groups.

6.2.2.1 Experimental item responses

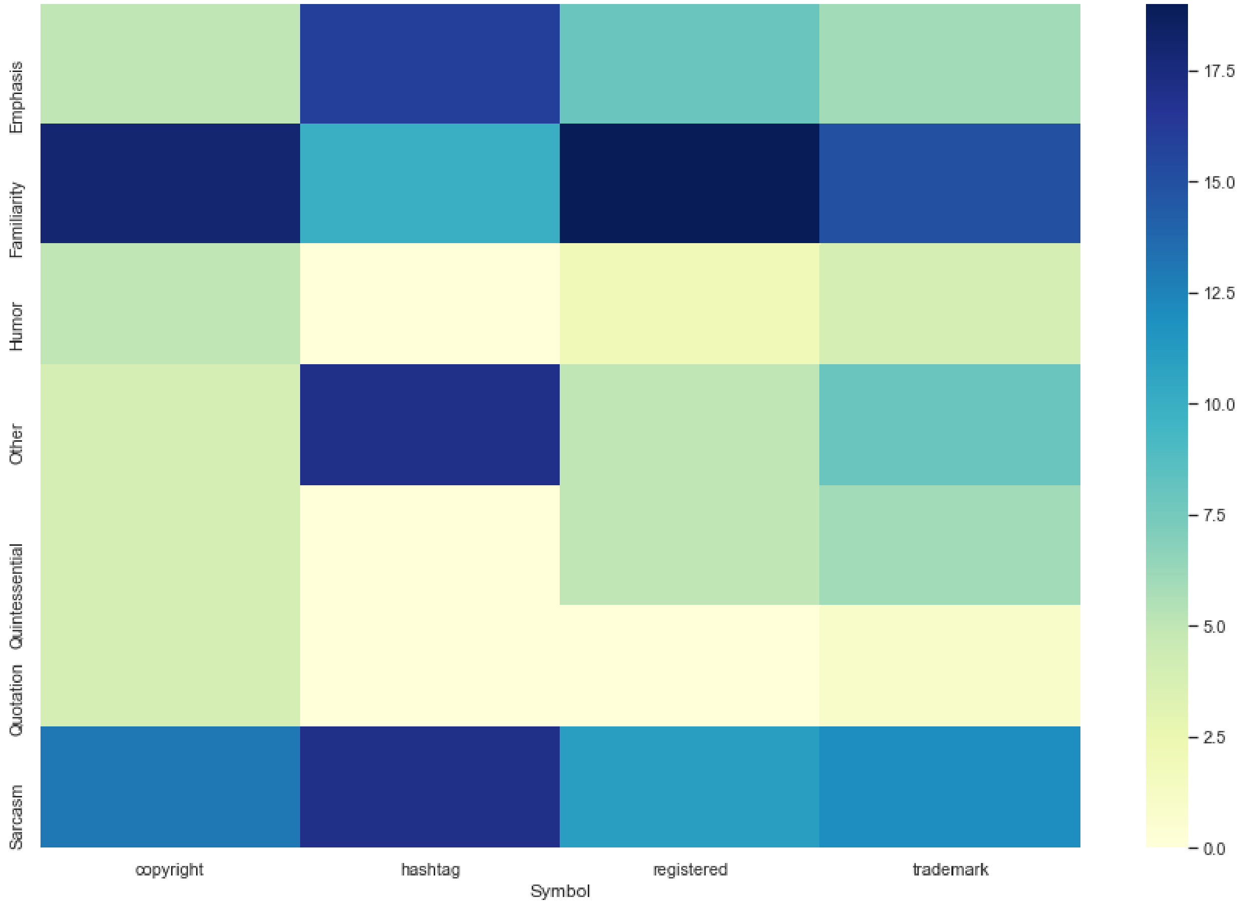
The distribution of responses for the experimental items is below in Figure 6.4.

Figure 6.4: Code distribution by operator for survey 1 experimental items



The literal category was the most popular across all 4 operators, suggesting that even in figurative uses, many people perceive them as performing their genuine function. However, users also posited a number of other functions, with emphasis, familiarity, and sarcasm all receiving more than 30 total responses. Looking at this distribution, there are also observable differences between the operators. For example, the hashtag received substantially more responses related to emphasis compared to the other operators. It also had far fewer responses related to familiarity compared with the other three operators. This difference is demonstrated again in the heatmap below in Figure 6.5.

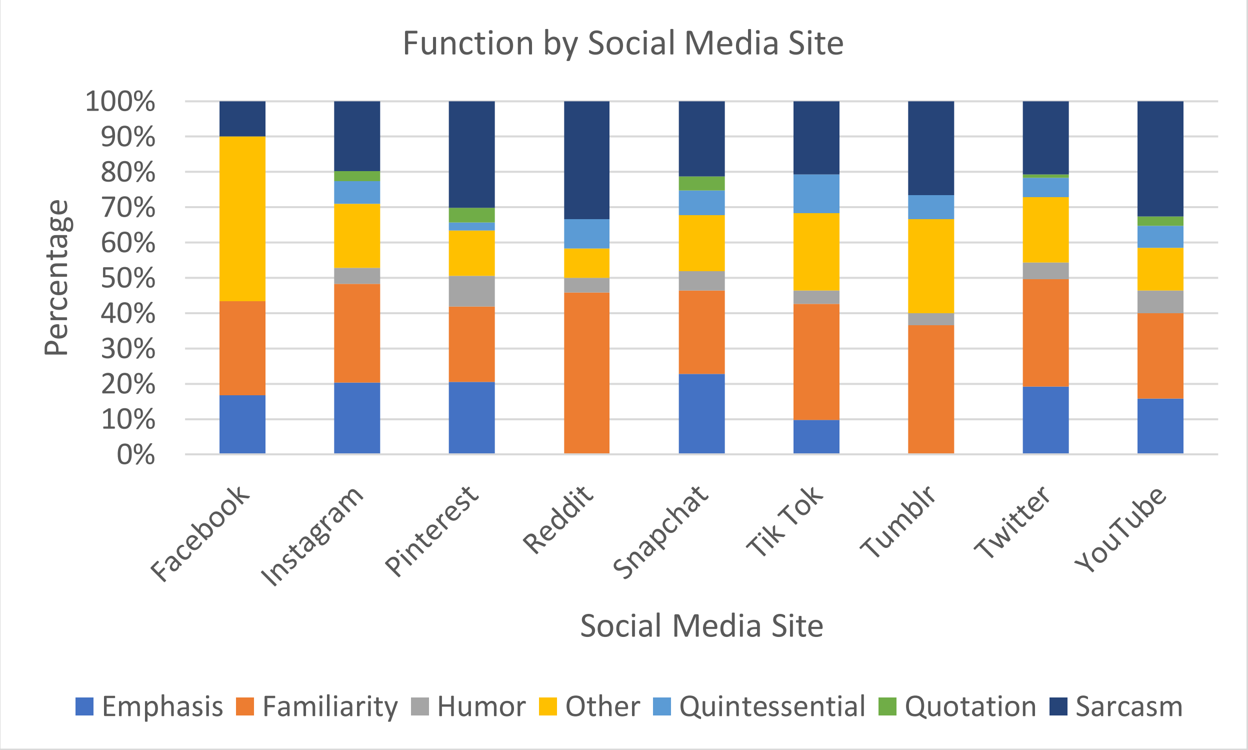
Figure 6.5: Heat map of functions by operator from survey 1



In the heatmap, it is easy to see that the hashtag received a higher response rate for sarcasm and emphasis than the other operators, and a lower response rate for familiarity. These differences were part of what motivated the design of Study 2. In Study 1, I did not control the items in such a way that allowed direct comparisons between the operators, however, I will revisit this question in the next section.

In order to understand the potential influence of social media, I looked at function selection by social media site. Figure 6.6 below reveals that Facebook users had the majority of *Other* coded responses, suggesting that perhaps Facebook users were less familiar with these operators and did not know how to explain them. For readability, the numeric labels are not included on the figure below, but they may be found in Appendix A.

Figure 6.6: Survey 1 function by social media site



In summary, analyzing the experimental item responses revealed that English social media users associate multiple functions with AOs beyond the literal functions. While some responses indicated participant confusion about the role of AOs, there were more than sufficient responses which provided thoughtful and insightful descriptions of the functions of the AOs. These functions were used in the design of Survey 2, in order to understand the degree to which new participants would find them plausible and accurate. However, for the purposes of Survey 1, the variety of pragmatic functions posited by the participants was an exciting result.

6.2.2.2 Control item responses

Responses to control items were coded using the approach described previously. Responses describing the control items offered insights into what users typically associate these operators with, as well as teasing apart what differentiates 'new' functions from the 'old' ones.

In this section, I provide examples of control item responses that were coded as *literal*. I spend time discussing what each operator is associated with to demonstrate the kinds of contrasts that I looked for when coding responses.

6.2.2.2.1 Genuine trademark

When participants received control items containing the trademark symbol, their responses coalesced around the legal features of the trademark, and its association with branding and companies. For example, one of the trademark control items was the following in example 156.

156. Thats not quite correct: theyve approved the Reb-A extract of Stevia, not the plant itself, but presumably **Truvia™** and **PureVia™** based sweetener packs should be available soon alongside Splenda and the rest.

Below are some example responses that this item elicited from participants.

157. To trademark Truvia

158. In order to show that the word is trademarked.

159. To show a registered item

160. Its showing the trademark on the brand name.

161. Truvia is a trademarked company. The author is making it clear that they are referring specifically to the official Truvia, and not another company/org./product by the same name.

Responses like the above were coded as *literal*. I interpret these responses as being related to the core, original purpose of the trademark. These explanations for the trademark included concepts like indicating an official product, brand or company. Specificity was also an important concept in many of the literal responses – that is, pointing to a specific instance of the trademarked entity.

However, not all control items were coded as *literal*. Even for the control items, people attributed the use of the operator to other motivations. For instance in the following example: *PB*

& Nutella™. Love Nutella! Goes great with PB, for sure While there were some answers that fell into the *literal* category, other participants attributed the purpose to humor.

162. the author jokingly used the trademark symbol to emphasize the serendipitous nature of Nutella and peanut butter. That, or there is a Nutella product that mixes Nutella and PB.

Additionally, for the control item: *That article really made me want to eat some* ***McDonalds™*** *fries.*, users brought up humor and emphasis.

163. Probably to be funny and mention the trade marked food place

164. To emphasize the type of fries

However, the majority (80%) of the control items for the trademark received responses that were coded as literal. It is interesting and exciting to note that these control items, which the researcher perceived as genuine, still received some descriptions and codes similar to the figurative ones. This might suggest that genuine instances of the operator may be used to achieve additional pragmatic effects, like humor and sarcasm. However, while the boundary between these categories may not always be clear, there was an overwhelming majority of responses for the control items which were related to the legal trademark, branding, and companies.

6.2.2.2.2 Genuine hashtag

The control items for the hashtag were generated such that it was clear the author was referring to an existing or trending hashtag. For example, in the excerpt below, the author refers to people who are posting using a specific hashtag.

165. I agree, and there are people who, I suppose are from other churches or other parts of the world, that are still posting using #wakeupolive to say that they're not giving up, she will rise, etc…

As with the TM, many participants responded with descriptions that were coded as *literal*. For the hashtag, literal codes had themes around movements, trends, and increasing the visibility of a post.

166. The highlighted symbol is being used as a hashtag.

167. they are referring to the category of posts that use and thus fall under the hashtag

168. To put the post under the antiinstragram hashtag, so other people using that hashtag will see their post.

169. To get the word trending

170. I believe the symbol was used to show that the author is part of the anti-Instagram group and have other people see the post.

171. That the word after is the name of a trend/movement

172. to reach more people

173. Used to get key words viewable and reused on social media

174. Symbol allows you to search Lisse.

175. That the letters behind it indicate some sort of movement/group/trend/etc

176. Trying to start a trend with the phrase following the symbol. Wants others to use the hashtag

As with the TM, the control items for the hashtag occasionally elicited more figurative responses, such as emphasis and familiarity.

Emphasis

177. To bring attention to the theme of the comment

178. To bring attention to a topic

179. To bring emphasis to something that has existing recognition.

Familiarity

180. That it is a commonly used saying, trend, movement used by people of that group

181. That this is the name of an artist/author that many people know about

However, the majority (79%) of the control HTs received responses that were coded as literal.

6.2.2.2.3 Genuine copyright

The control items for copyright involved crediting a source, legal copyright, and plagiarism..

182. It is used to show that the source has copyrighted its material.

183. To give notice of copyright

184. In order to state that the user got the information from an official company to state their point of view.

185. To indicate that this excerpt is protected on a legal basis from being plagiarized

186. Because the ESV Bible is seen as an original piece of work and is being quoted in the comment.

187. Showing the copyright year.

188. Normal copyrite usage.

189. They cited an article and wanted to let the audience know when the research was released

190. To give credibility with the recognizable symbol

191. To specify that the information came from an official article

192. intended use of symbol

193. To shows the source of where the excerpt comes from.

194. To show when the comment was made and to add credibility

However, one interesting occurrence was that some users felt that the copyright symbol was functioning as punctuation.

195. This is used as a bullet point.

196. This symbol looks like its being used as a comma.

Compared with the previous two categories, an even higher percentage of items were coded as *literal* for the copyright symbol, with 86.7% being coded as the literal function. Unlike with the hashtag and trademark, there were no responses in the control items for copyright that fit into any of the figurative codes.

6.2.2.2.4 Genuine registered trademark

Responses that were coded as literal for the registered trademark shared overlap with both the trademark and copyright symbols. Registered trademark was associated with branding, companies, slogans, as well as copyright and legal protection of content.

197. rights reserved like its just the company wording stuff

198. For official business purposes as it certifies a product as officially produced by Microsoft.

199. Specific brand

200. To correctly trademark a company

201. The highlighted symbol indicates that LEGO is a registered trademark.

202. Represents Lego as a company.

203. To inform that Lego is a brand

204. That it is a slogan that has been copyrighted to prevent illegal copying

205. To give credit to the software used, and make it clear that they used the official photoshop software from Adobe and not another brand's.

206. This symbol has something to do with legalities.

207. That it is the name of a product/company name that cannot be copied and is well known

208. Because its for a word of a product a company owns

However, some participants associated the registered trademark with a desire to sound *fancy* or *more educated* as in the examples below.

209. reserved? like to be fancy and extra. who has the time to add symbols

210. That symbol marks a full trademark. They might have used this instead of the plural LEGO’s like the rest of us would have to sound fancy or more educated?? I don’t know.

The first response notes the extra time and effort required to add the symbol, suggesting that most people would not have *the time to add symbols.*

In summary, participants demonstrated a robust understanding of the genuine functions of AOs. For the control items, they provided responses that were overwhelmingly related to those genuine functions. Responses also indicate that the boundary between these operators is not always clear, but the goal of this section was to demonstrate that the responses from participants provided clear contrasts between *literal* and *figurative* meaning.

6.2.3 Statistical analysis

Figure 6.7: Survey 1 response codes for control items.

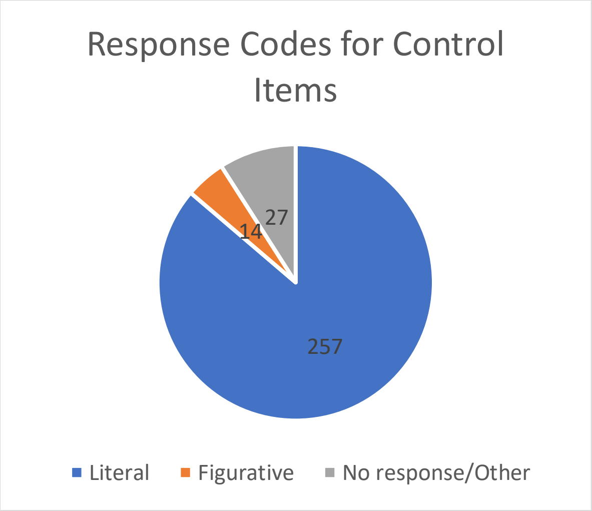
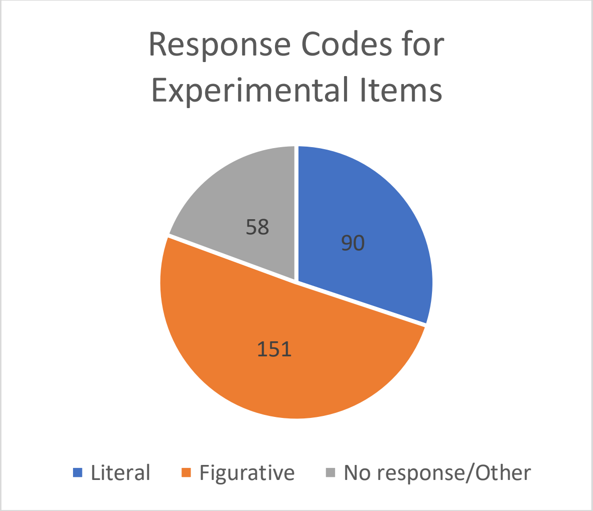


Figure 6.8: Survey 1 response codes for experimental items



After the initial coding of responses, responses were re-coded in a simplified manner into one of three categories: *Literal*, *Figurative*, and *No Response/Other*. There were some items which received both *Literal* and *Figurative* codes. For the statistical analysis, these items were treated as *Literal*. For the control items, 257 responses were coded as Literal, 14 as Figurative and 27 as *No response/Other*. *No Response/Other* codes were filtered out from the dataset before statistical testing. For the control items, this left a total of 271 responses collected. Examining the raw frequencies, there is a large difference between the Literal and Figurative groups, with the control items unsurprisingly having a majority coded as *Literal*. A chi-squared goodness of fit test (X-squared(1) = 217.89, p<.001) demonstrated that the response frequencies are significantly different from one another. In other words, for the control items, there were significantly more *Literal* responses than *Figurative* responses.

For experimental items, after filtering out *No Response* codes, there were 241 responses, 151 Figurative and 90 Literal. A chi-squared goodness of fit test (X-squared(1) = 15.44, p<.001) demonstrated that the response frequencies for the experimental items were also significantly different from one another, with significantly more responses coded as *Figurative* compared to *Literal*. Although this difference is not as dramatic as it was in the control items, there was still a statistically significant preference for figuratively coded responses for the experimental items.

Finally, a chi-squared test for independence demonstrated that the observed distribution of *Literal* and *Figurative* codes differs reliably from the expected distribution if there was no relationship between item type (control vs. experimental) and response code (literal vs. figurative) (X-squared(1) = 190.4, p < .001). In other words, there is evidence that there is a significant difference between participant perception of *Genuine* uses of AOs and *Figurative* uses of AOs, with participants being reliably more likely to describe the control items as serving the default or literal meaning of the operator, and more likely to describe the experimental items as serving a figurative purpose.

6.3 Survey 2: between operators

6.3.1 Design and participants

Survey 2 was designed leveraging the results of Survey 1. The main six figurative functions that participants described in Survey 1 became responses for the items in the following study. Survey 2 was intended to build upon these results by seeing if these functions would resonate with a new group of participants. Additionally, Survey 2 was designed with the goal to explore whether under controlled conditions the operators would be perceived as contributing different meanings, and whether or not particular operators are more likely to be associated with particular functions.

As with Survey 1, all the items presented excerpts from Reddit. Excerpts were chosen which be reasonably short (~1 paragraph or less), as opposed to long, multi-paragraph posts. Additionally, excerpts were chosen such that each operator had the same number of items (3 per operator). Items were rated as either short (<10 words), medium (10-20 words), or long (>20 words). For each operator, 3 examples were chosen: one short, one medium, and one long. In order to perform comparisons across the operators, four different survey conditions were created. Each condition had three items for each operator, creating a total of 12 items. The conditions varied according to which operator was paired with the affixed unit. The full list of items may be found in Appendix D. An example of one item in all four conditions is in Table 6.2 below.

Table 6.2: Example item variation across four conditions for Survey 2

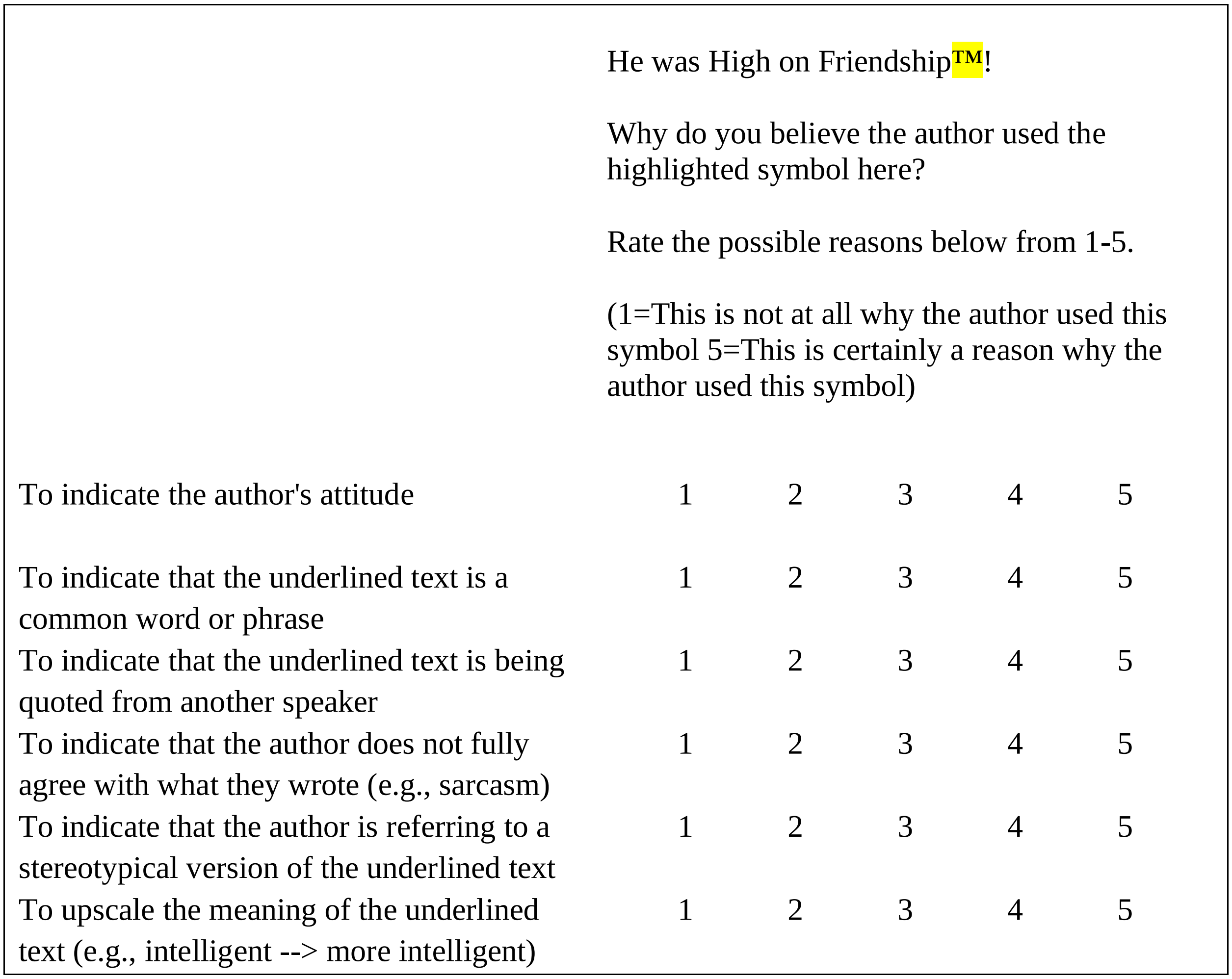
Version A	Version B	Version C	Version D
He was High on Friendship™!	He was High on Friendship©!	He was High on Friendship®!	He was #HighonFriendship

In the case of this item, Version A (*He was High on Friendship*™) was the original post from Reddit, with the rest of the versions containing a modified post. The most overt modification is in the case of the hashtag, which comes in front of the affixed unit, and additionally requires that the words occur with no spaces in between. However, as these are normal conventions for the hashtag, it was seen as necessary to follow this format in order to simulate a naturally occurring hashtag.

Items were balanced to ensure that every condition contained a mix of original and modified posts. However, balancing the number of items per AOs was prioritized. It was mathematically impossible to balance both AOs and original items, so in the end, two conditions contained 3 original items, one condition contained 4 original items, and one condition contained 2 original items.

When presented with an item, participants were asked to rate different possible functions. The functions were *attitude*, *familiarity*, *quotation*, *sarcasm*, *stereotype*, and *upscaling*. The functions were determined by the results of the previous study, alongside insights from the qualitative analysis. For example, emphasis became upscaling. Humor was exchanged for author attitude, in light of the important role that evaluative meaning was seen to perform in the qualitative analysis. A sample item is presented below in Figure 6.9. The full list of items may be found in Appendix D.

Figure 6.9: Survey 2 example item



He was High on Friendship™!

Why do you believe the author used the highlighted symbol here?

Rate the possible reasons below from 1-5.

(1=This is not at all why the author used this symbol 5=This is certainly a reason why the author used this symbol)

To indicate the author's attitude	1	2	3	4	5
To indicate that the underlined text is a common word or phrase	1	2	3	4	5
To indicate that the underlined text is being quoted from another speaker	1	2	3	4	5
To indicate that the author does not fully agree with what they wrote (e.g., sarcasm)	1	2	3	4	5
To indicate that the author is referring to a stereotypical version of the underlined text	1	2	3	4	5
To upscale the meaning of the underlined text (e.g., intelligent --> more intelligent)	1	2	3	4	5

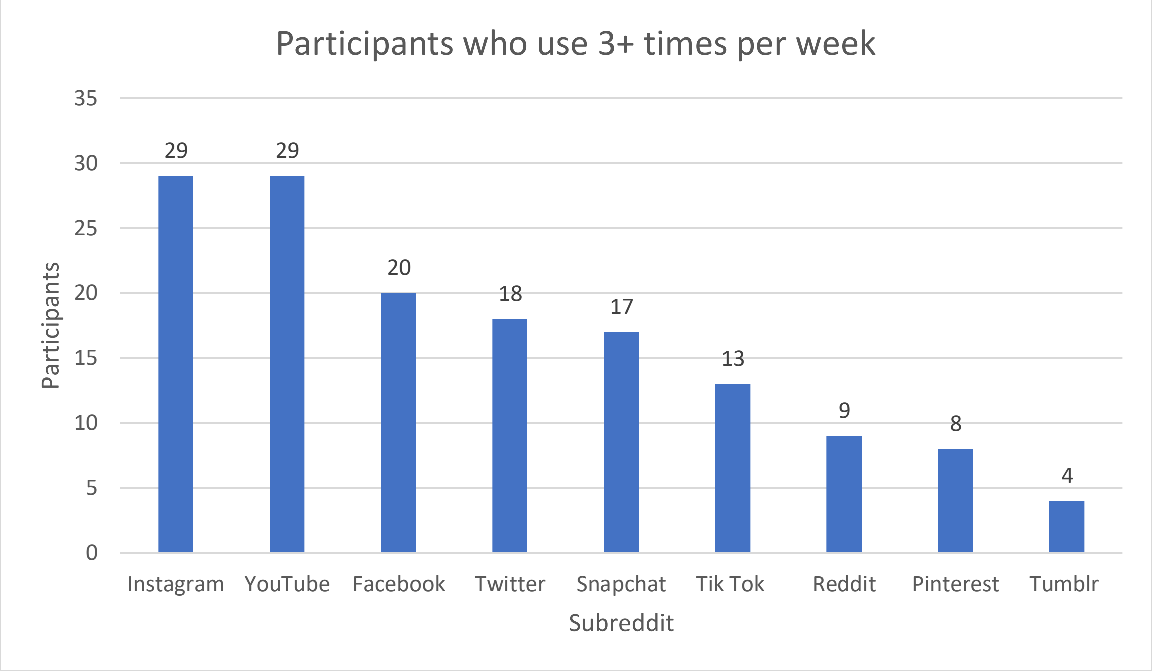
Each participant saw 12 items and performed 6 functional ratings for each item. Participants were also asked a series of questions about their background before they began the study, including their social media use.

40 participants were recruited to participate in the study, 10 for each condition. Participants were recruited in two ways. Undergraduate students enrolled in Linguistics courses at the University of Texas at Arlington were offered extra credit to participate in the study. Additionally, acquaintances of the researcher were recruited via email. 16 men and 24 women participated. 12 participants reported being between the ages of 18-24, 9 between 25-29, 9

between 30-39, 4 between 40-49, 5 between 50-59, and 2 being 60+. Notably, compared to Survey 1, a much broader range of ages participated. 35 participants were L1 users of English, 5 were L2 users of English. Each of the L2 users of English reported having more than 5 years of experience speaking English.

As in Survey 1, participants were asked to report which social media platforms they use 3+ times per week. The distribution of people per social media site is reported in Figure 6.10 below.

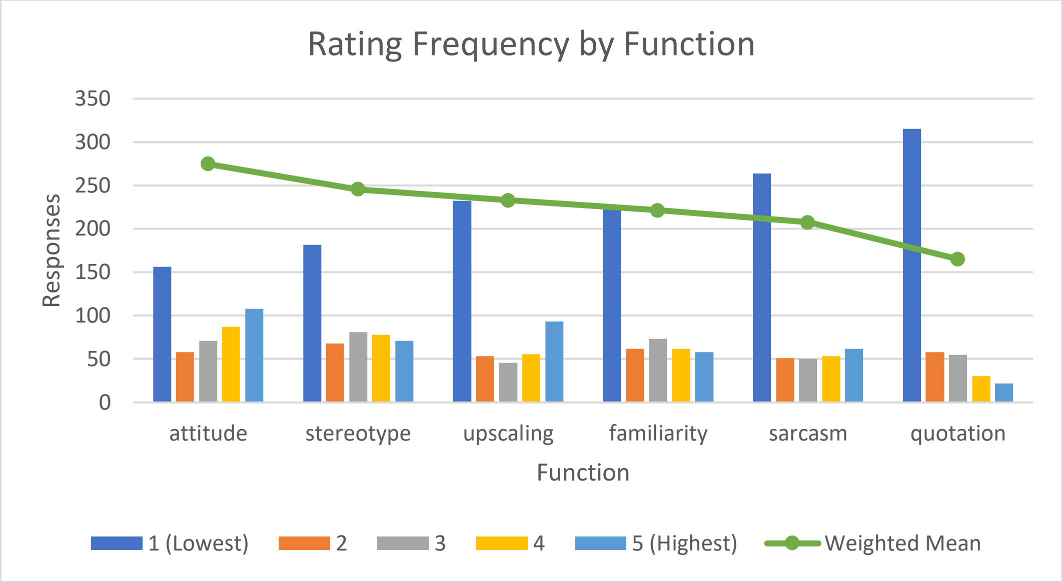
Figure 6.10: Survey 2 participant social media use



6.3.2 Descriptive analysis

Each participant rated 12 items for 6 functions, yielding 72 observations per participant. This generated a total of 2880 ratings.

Figure 6.11: Rating frequency by function



In Figure 6.11 above, it is clear that *1* was the most frequent rating across the board. This is perhaps unsurprising, as for any given function, some portion of the examples would not be an instance of that function and would therefore receive mostly low scores. In this way, every function is expected to have some portion of *1* responses. However, it is interesting to note those functions which had a majority of the responses as *1*, such as quotation, which was clearly a function that participants believed to be less relevant. The Weighted Mean line represents the total sum of the values for each function divided by the 5 possible values. As participants were asked to rate the extent to which they believed a function applied to a particular example, the higher the weighted mean, the higher overall values assigned to that particular function across the board. In aggregate, attitude received the highest scores, with quotation receiving the lowest.

Figure 6.12: Survey 2 rating frequency by function (TM)

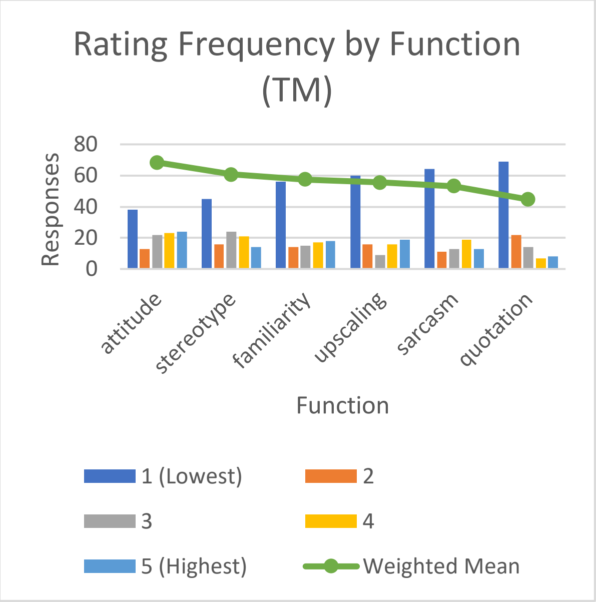


Figure 6.13: Survey 2 rating frequency by function (HT)

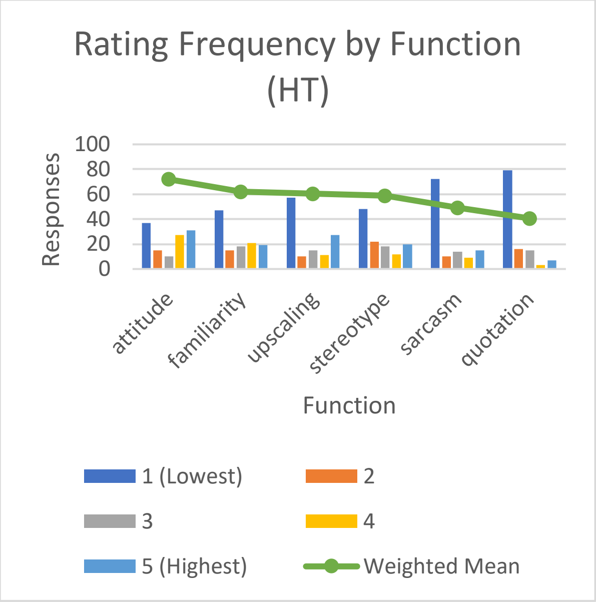


Figure 6.14: Survey 2 rating frequency by function (R)

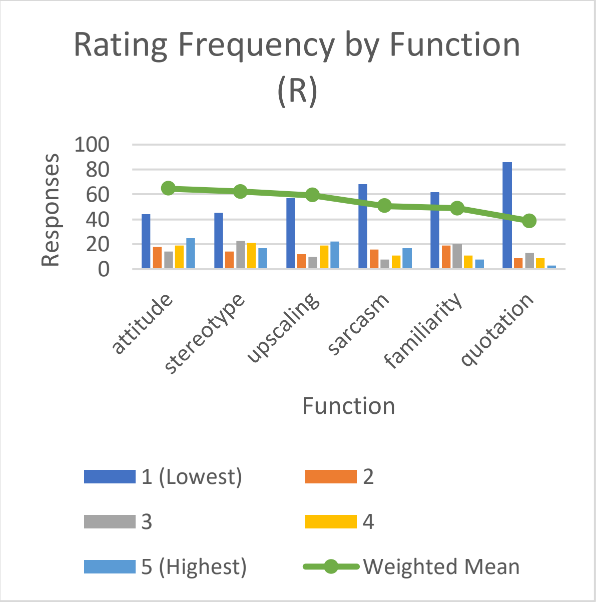
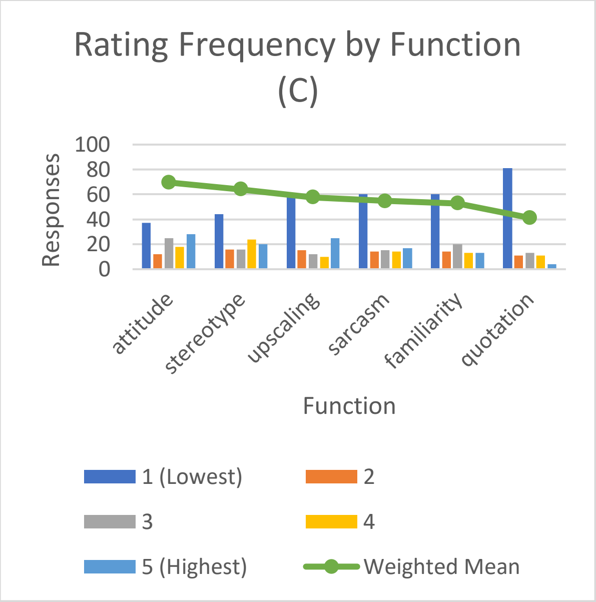


Figure 6.15: Survey 2 rating frequency by function (C)



When broken out by operator, the overall trends of attitude being the top and quotation being the lowest were the same. However there were some differences among the operators in terms of the weighted averages and in terms of the internal ordering. The registered trademark (®) and copyright symbol (©) had an identical order of functions: *attitude*, *stereotype*, *upscaling*, *sarcasm*, *familiarity*, and then *quotation*. The trademark (™) had a similar, but slightly different order with *attitude*, *stereotype*, *familiarity*, *upscaling*, *sarcasm*, *quotation*. Finally, the hashtag (#) order was *attitude, familiarity*, *upscaling*, *stereotype*, *sarcasm* and *quotation*. Notably, familiarity scored higher for the trademark and the hashtag, while sarcasm scored higher for the copyright and registered trademark.

The main hypothesis being tested in this study was whether the operators would be perceived as contributing different meanings when paired with the exact same text. Because participants input values that were scalar in nature, Ordinal Logistic Regression (OLR) was selected as an appropriate model to fit to the data. As with other approaches to regression, in OLR there is one dependent variable, in this case the rating between 1-5, and one or more independent variables. In this case, the independent variables were operator type, function, item, and participant. OLR differs from ordinary multinomial regression in that it takes into account how the different responses are ordered. In summary, an OLR effectively allowed for an understanding of whether or not there is any meaningful relationship between operator, function and value. The ordinal logistic regression models were then used to calculate the estimated marginal means using the r package emmeans (Lenth, 2019) to allow for pairwise comparisons.

6.3.3 Ordinal logistic regression modeling

A total of ten mixed-effects Ordinal Logistic Regression (OLR) models were fit to the data. The dependent variable for all 10 models was the value of the rating (1-5) that participants gave for a particular function for a particular item. For each function model, the independent variable was operator. For each operator model, the independent variable was function. For each model, the item and subject were treated as random effects.

For each OLR model, the R package emmeans was used to do pairwise comparisons. For the function models, the pairwise comparisons were done between the 4 operators. That is, the model was examining if for a given function, there were significant differences between the operators. The operator models were the inverse: for each operator, the pairwise comparisons were done between the functions. In summary, the statistical analysis was designed around questions about the relationship between function and operator. Key findings from the models are summarized in the section below. The full results from the OLR models and emmeans pairwise comparisons may be found in Appendix E.

6.3.3.1 Operator models

6.3.3.1.1 Trademark model

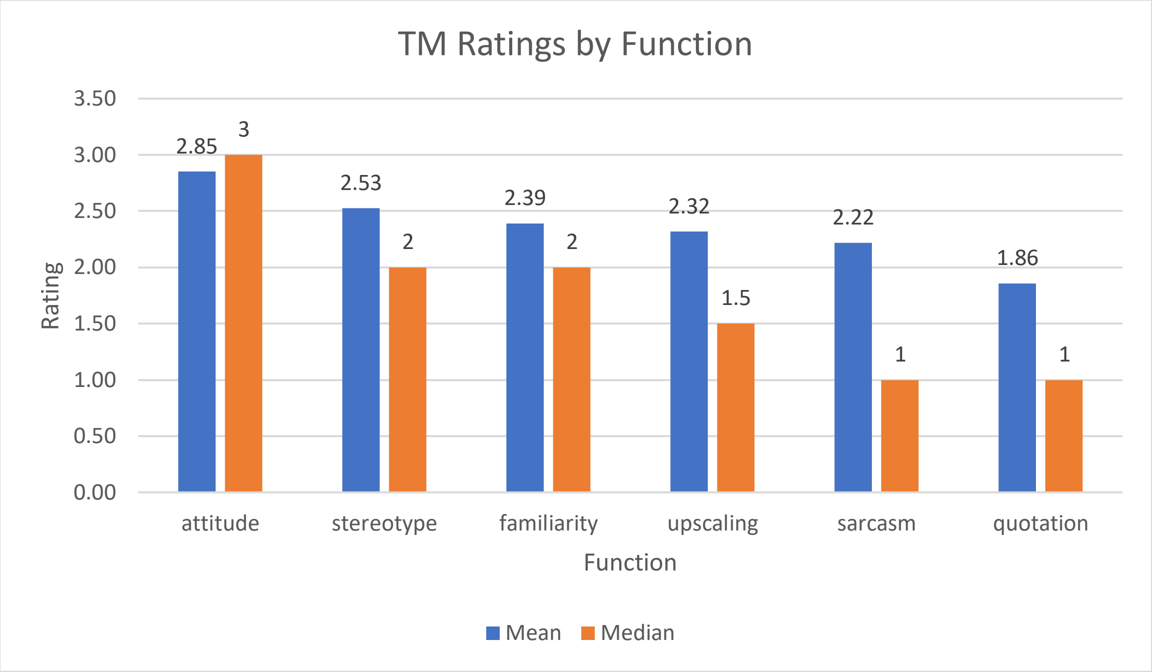
The trademark model had three significant features, below in Table 6.3.

Table 6.3: Significant features from the trademark model

Feature	**Coefficient**	**Significance**
Function: Quotation	-1.8009	0.00195
Function: Sarcasm	-1.1680	0.01885
Function: Upscaling	-0.9517	0.01272

All the functions had negative coefficients, with quotation being the most negative and the most significant. Pairwise comparisons were done between the six functions using emmeans. Attitude and quotation ($p<0.05$) were significantly different for the trademark model. Mean rating for attitude for the trademark symbol was *2.85* compared to *1.86* for quotation. The mean trademark ratings for all the functions are below in Figure 6.16.

Figure 6.16: Survey 2 trademark ratings by function



6.3.3.1.2 Hashtag model

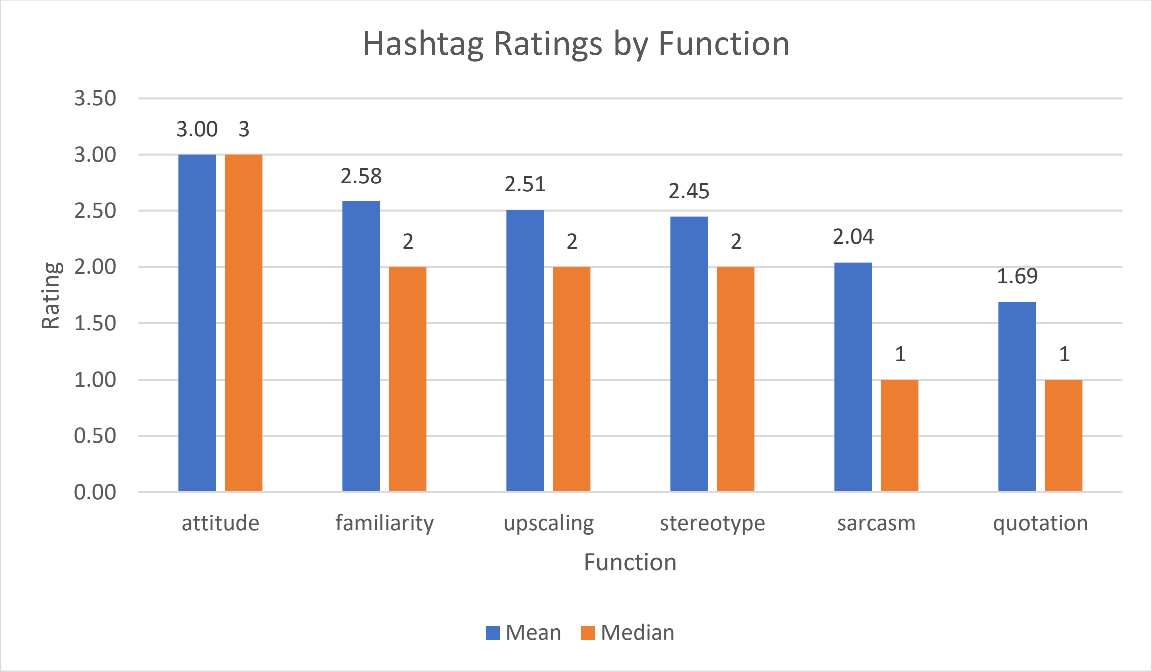
The hashtag model had two significant features.

Table 6.4: Significant features from the hashtag model

Feature	Coefficient	Significance
Function: Quotation	-2.5249	3.30E-05
Function: Sarcasm	-1.8168	0.00559

Both functions have negative coefficients, and as with the TM model, quotation and sarcasm are the two most negative and significant functions. Pairwise comparisons were done between the six functions using emmeans. Attitude and quotation (p<0.001), familiarity and quotation (p<.01), and quotation and stereotype (p<0.05) were significantly different for the hashtag model. Mean rating for attitude for the hashtag symbol was *3* compared to *2.58* for familiarity, *2.04* for sarcasm and *1.69* for quotation. The mean hashtag ratings for all the functions are below in Figure 6.17.

Figure 6.17: Survey 2 hashtag ratings by function



6.3.3.1.3 Registered trademark model

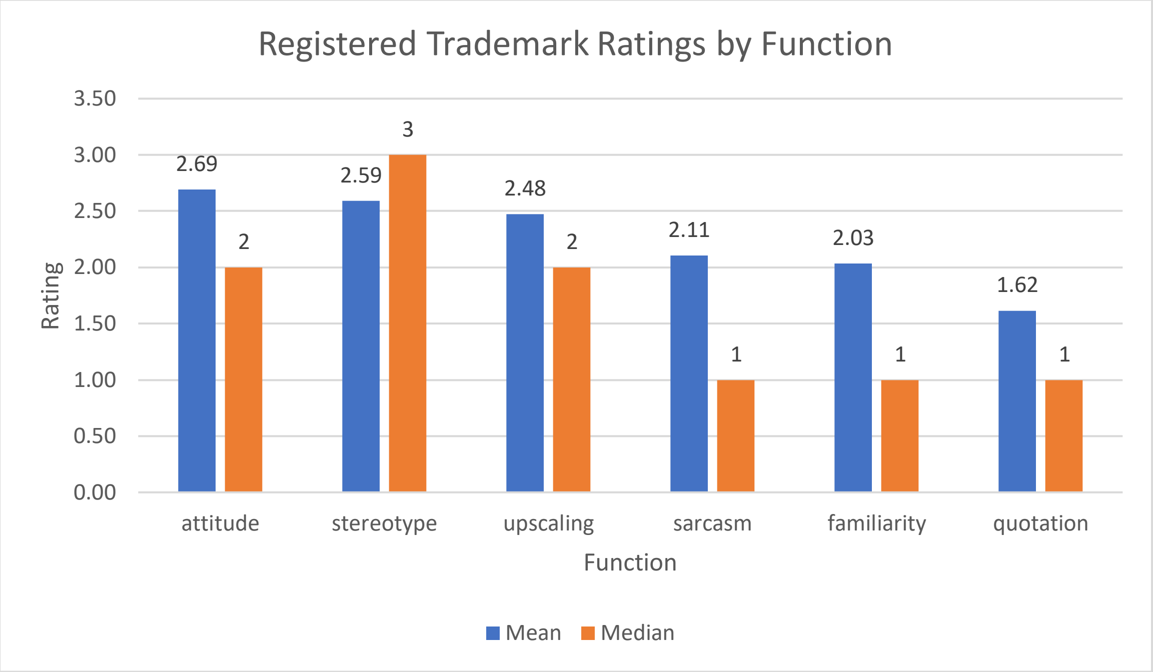
The registered trademark model had three significant features.

Table 6.5: Significant features from the registered trademark model

Feature	Coefficient	Significance
Function: Quotation	-0.9685	0.015178
Function: Sarcasm	-2.1756	0.000135
Function: Familiarity	-1.0842	0.021441

All three functions have negative coefficients. Pairwise comparisons were done using emmeans. Attitude and quotation (p<0.005), and stereotype and quotation (p<0.001) were significantly different for the trademark model. Mean rating for attitude for the registered trademark symbol was *2.69* compared to *2.11* for sarcasm and *1.62* for quotation. The mean registered trademark ratings for all the functions are below in Figure 6.18.

Figure 6.18: Survey 2 registered trademark ratings by function



6.3.3.1.4 Copyright model

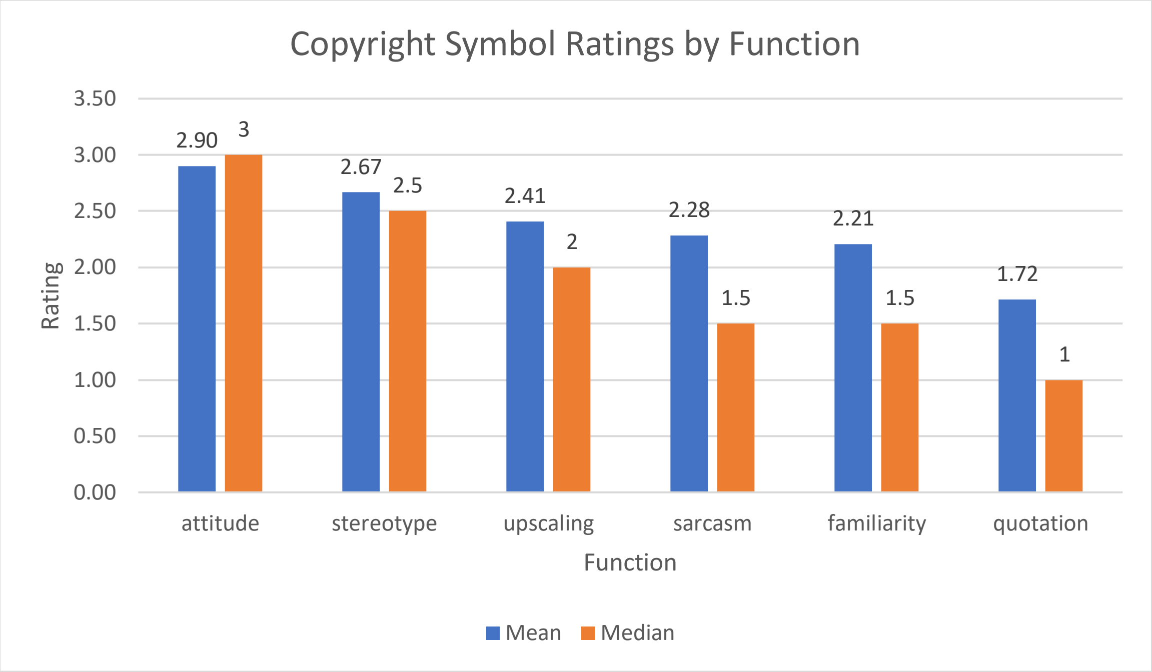
The copyright symbol model had three significant features.

Table 6.6: Significant features from the copyright model

Feature	Coefficient	Significance
Function: Quotation	-2.101	0.0000227
Function: Sarcasm	-0.9213	0.0454
Function: Familiarity	-1.0357	0.0189

As with all the other operator models, quotation and sarcasm are significant and have negative coefficients. Additionally, the copyright symbol patterns after the registered trademark symbol in having familiarity be significant and negative. Pairwise comparisons were done using emmeans. Attitude and quotation (p<0.001), and stereotype and quotation (p<0.001) were significantly different for the copyright model. Mean rating for attitude for the copyright symbol was *2.90* compared to *2.67* for sarcasm and *1.72* for quotation. The mean copyright ratings for all the functions are below in Figure 6.19.

Figure 6.19: Survey 2 copyright symbol ratings by function



6.3.3.2 Function models

Ordinal logistic regression models were fit for each of the 6 functions. For each model, *item x symbol* and *subject x symbol* were passed as random effects. None of the six function models yielded significant features. Each of the models was passed to emmeans to do pairwise comparisons. None of the pairwise comparisons showed significant differences between the operators.

6.4 Conclusion

In this chapter I have detailed the results of two surveys on AOs that I conducted with social media users. In Survey 1, participants were presented with excerpts from Reddit and gave free input responses describing the motivation behind the use of the AO. Participants saw control items, which were instances of genuine AOs, and experimental items, which were instances of figurative AOs. Participant responses were coded into nine functional categories: *emphasis*, *familiarity*, *humor*, *quintessential*, *quotation*, *sarcasm*, *literal*, *other*, and *no response*.

One goal of the study was to corroborate the qualitative analysis. Many of the functions described here do indeed correspond to the functions I identified in chapter 4 and chapter 5. I argued that sarcasm (distancing), familiarity marking, and quintessence (prototypicality marking) are manifestations of stance marking, and each of these functions was present in the participant responses. I furthermore have noted that AOs may upscale the meaning of the affixed units. While participants did not explicitly mention this function, the participant descriptions of AO *emphasis* have some overlap with the kind of scalar modification we saw in chapter 5.

With the control items, participants described genuine functions of AOs. These descriptions revealed differences between the way participants perceived genuine AOs and figurative AOs. Importantly, these descriptions also revealed differences between the way participants perceived the genuine meanings among the four operators. Hashtags were associated with findability, copyrights were associated with protecting original works, trademarks were associated with slogans and products, and registered trademarks appeared to have a combination of copyright and trademark genuine functions. The differences among the genuine functions of the AOs are particularly interesting, because we do not see the same degree of differences between the figurative operators.

The results of a chi-squared test for independence revealed that participants were more likely to label figurative AOs with a figurative function and genuine AOs with a literal function. In other words, there does appear to be some use in distinguishing between these two categories, as participants perceived them differently. However, close examination of participant responses also revealed that some participants associated genuine AOs with figurative functions and figurative AOs with literal functions. This provides evidence that the boundary between these categories is not always clear. This is an exciting finding because it suggests that even for figurative AOs, the genuine meaning is likely still activated in the context. It seems plausible then, that these functions are both be contextually present when people leverage AOs, whether figurative or genuine.

In Survey 2, participants were presented with an excerpt from Reddit and asked to rate the accuracy of six different functional descriptions for the AOs: *attitude, familiarity, stereotype, quotation, sarcasm* and *upscaling*. Participants selected the evaluative function (*attitude*) of the operator most often and with the most confidence, with author attitude being associated with a score of 4 or 5 in 195 observations out of 480 observations related to attitude (40.6%). In other words, in nearly half of the examples, participants agreed that the operator contributed evaluative meaning.

Ordinal logistic regression models were fitted to the data with a total of 10 models—one for each function and operator. In the operator models, the fitted models had significant features, and showed some significant differences in the emmeans output. All 4 of the operator models had quotation and sarcasm as significant predictors with negative coefficients, suggesting a broad pattern of users rating these functions lower. All 4 of the distinct operator models showed significant differences between the attitude function and the quotation function, with attitude receiving significantly higher values than quotation. A few insights can be extracted from this. First, participants did not provide strong evidence for quotation as a function of AOs. Second, participants were much more likely to rate attitude highly as a function of AOs.

The fact that none of the function models revealed significant differences is perhaps unsurprising. This corroborates the findings of previous chapters which reveal the operators behaving in similar ways. While we cannot ‘prove’ a lack of differences between the operators with these results, the output of the function models, at the very least, do not point to significant differences between the operators in controlled conditions. That is, context appears to be a bigger driver of functional meaning, which explains why in the operator models, we saw similar functions showing up as significant across all the operators.

Lastly, I would like to note that familiarity was a significant feature in every model except for the hashtag model. While for the copyright and registered trademark, familiarity had the second lowest mean score, for the hashtag and trademark, familiarity had the second and third highest mean scores respectively. Therefore, it seems like familiarity is one function where the operators may differ. Although the familiarity model itself yielded no significant results in the pairwise comparisons, it did show a difference approaching significance between the hashtag and registered trademark (p=0.0641). Therefore, we might derive an insight that the hashtag was associated with familiarity marking for participants, whereas the copyright and registered trademark were not.

CHAPTER 7

INSIGHTS FROM MACHINE LEARNING

In this chapter, I overview the results from applying machine learning and data mining approaches to the Pushshift Reddit dataset. This approach was motivated by the massive size of the corpus. With ~5M comments containing AOs, it was not within the scope of this project to examine each data point, or even each subreddit, individually. Therefore, in addition to the corpus linguistics techniques I outlined in chapter 3, I considered machine learning a valuable approach to derive insights about the entire corpus that might not otherwise be detectable.

While the corpus linguistics analysis in chapter 3 focused primarily on describing features of AOs themselves, in this chapter I turn to the authors who use AOs. User behavior was selected as an achievable, quantifiable aspect to model, while also having relevance to the broader goals of this work. The goal of these models was to predict AO use at the individual author level to provide insight into the importance of features such as community membership, tenure within a community, and tenure on Reddit in predicting AO use. As such, these models were built to address the research questions: *Can we predict AO use on Reddit with machine learning models? Will author-related features or subreddit-related features be more important to the models?*

User behavior was explored in two ways. The first was attempting to quantify the volume of an author's AO use. For this model, the goal was to take a particular author and predict, for each of the subreddits they have ever posted in, how many AOs they posted in that subreddit. This became a regression task, as the target—the number of AOs posted in a subreddit—is continuous, not discrete. The second way I explored user behavior was by predicting whether a particular author has ever used an AO. For this model, the goal was to take a particular author’s post history and predict whether they are an AO user. This became a classification task, as the target that the model was predicting was binary: either yes or no.

Machine learning algorithms range in their degrees of interpretability, but for my purposes, a highly interpretable ML algorithm is one which provides insights into which features are the most important for model success. For the regression task, linear regression and random forest regression were selected due to their ability to produce feature importances. Similarly, for the classification task, logistic regression and random forest classification were selected for their ability to produce feature importances. For both the regression task and the classification task, I used two algorithms: one that is simple, well-founded, and well-understood—linear and logistic regression—and one that is able to capture more complex trends and potentially be a better fit to the data—random forest. Models such as neural networks and other large language models like BERT (Devlin et al., 2018) were not used because they fail to satisfy the interpretability requirement for this kind of work.

In this chapter, I compare the performance of various models, noting the role of different features and how they relate to model performance. Rather than prioritizing a model that can most effectively predict an author’s use of AOs, the goal was to build models that would yield insights into what features are most valuable and meaningful in predicting AO-usage. In the remainder of this chapter, I detail the datasets, features, and results of the models for these two ML tasks. I also discuss the feature importance results to derive insights that link back to the broader goals of this dissertation.

7.1 Dataset

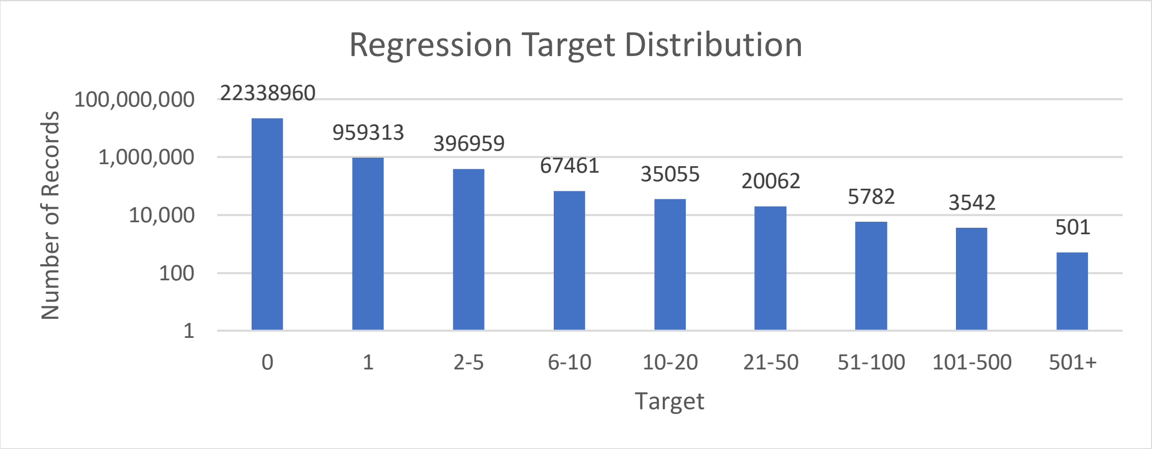
The dataset for both the regression and classification tasks contains aggregated information from the broader Pushshift Reddit dataset (Baumgartner et al., 2020) which was, in this case, downloaded via pushshift.io rather than accessed via Google Big Query (Fernandes & Bernardino, 2015). The dataset hosted on pushshift.io spans from December 2005 to June 2021, containing 18 more months than the Big Query copy which was analyzed in the corpus analysis. While in the corpus analysis, there was a need to perform dynamic queries of the data, for the machine learning, a single one-time snapshot was suitable. Therefore, I switched to the Pushshift version of the dataset for this task, in order to take advantage of the additional 18 months of data. Data wrangling and transformation was generously performed by a technical consultant from University Analytics at UT Arlington. Because the aim of these analyses was a large-scale data mining approach based in machine learning, this dataset underwent a different cleaning process than the data I analyzed in the corpus linguistics work of chapter 3. Specifically, the fasttext-langdetect Python library was used to filter out non-English posts based on its balance of speed (which is critical for analyzing such a large corpus) and accuracy against benchmark datasets. Fasttext-langdetect is a Python wrapper around Facebook's FastText-based models (Joulin et al., 2016) for language identification. Only posts with a score of 0.95 or higher according to the fasttext-langdetect library were kept in the dataset. After filtering, the corpus contained a total of ~5 billion Reddit comments and ~200 billion words.

7.2 Regression task

The first Machine Learning task was a regression problem which involved predicting how many AO-containing posts an author made in a given subreddit. Of particular interest is whether features relating to the author (e.g., first post on Reddit) or the subreddit (e.g., frequency of hashtags) would be given more weight by the models. Specific features are discussed in more detail in section 7.2.1.

The dataset for the regression model was filtered such that it includes only authors who have used at least one non-hashtag AO and only subreddits that have at some point contained at least one non-hashtag AO. The emphasis on non-hashtags was due to the significant frequency disparity among the AOs, which I have detailed in chapter 3. If the dataset included authors and subreddits that have used *any* AO, the frequency disparity would have led to an imbalanced dataset, in which most of the datapoints only involved the hashtag. Therefore, to enable an understanding of what drives AO-usage across all symbols, the dataset included only those individuals and communities where at least one low-frequency AO post (™, ©, ®) was associated with them. This yielded a dataset with 269,419 unique authors, 21,502 unique subreddits and a total of 23,827,635 observations. The dataset contained one row per author per subreddit, excluding subreddits where the author made no posts. The target the model was aiming to predict for each row is the number of posts containing AOs that the author has made in a specific subreddit. The distribution of the target is strongly negatively skewed, with the majority of observations having a target of 0. The distribution of the target is below in Figure 7.1. Figure 7.1 and the other plots in this chapter have y-axes on a logarithmic scale unless otherwise noted.

Figure 7.1: Regression task target distribution



7.2.1 Feature generation

Feature generation and selection is an important step in any machine learning task, as model performance is often directly tied to the quality of the features. This is particularly crucial for a task like this, where understanding and interpreting feature importance is the ultimate goal. Therefore, features were generated with interpretability and relevance to the project goals in mind. Features were generated at an author level and a subreddit level. A description of each of the generated features is in Table 7.1 below. Each feature was also assigned to one of three feature sets: author, subreddit, or cross (related to both author and subreddit).

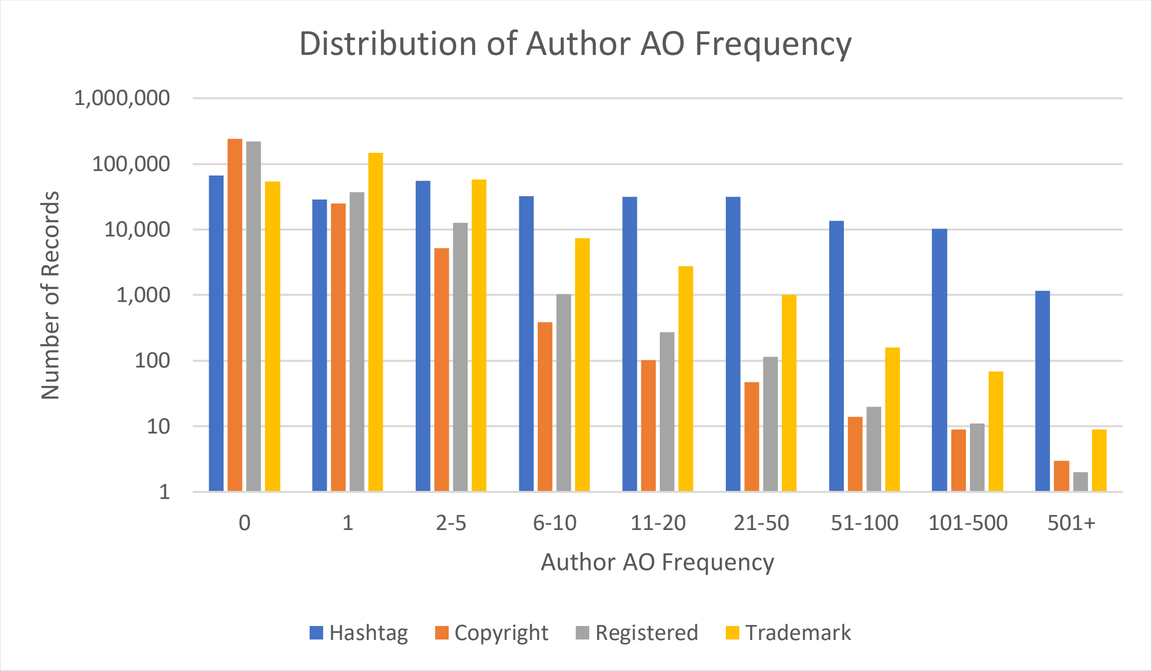
Table 7.1: Regression task features

Feature Set	**Feature**	**Description**
author	authors_hash/c/tm/r_total	Total number of hashtags/copyrights/trademarks/registered in author history
author	authors_hash/c/tm/r_comments	Total number of comments containing hashtags/copyrights/trademarks/registered in author history
author	author_word_count	Total number of tokens in author history
author	author_num_posts	Total number of posts in author history
author	author_reddit_min_date	Day, Month, Year of author's first post on Reddit
author	author_reddit_max_date	Day, Month, Year of author's latest post on Reddit
cross	author_posts_in_subreddit	The total number of times a given author has posted in a given subreddit
cross	author_subreddit_min_date	Day, Month, Year of first post by author in subreddit
cross	author_subreddit_max_date	Day, Month, Year of last post by author in subreddit
subreddit	subreddit_hash/c/tm/r_total	Total number of hashtags/copyrights/trademarks/registered in subreddit history
subreddit	subreddit_hash/c/tm/r_comments	Total number of comments containing hashtags/copyrights/trademarks/registered in subreddit history
subreddit	subreddit_word_count	Total number of tokens in subreddit history
subreddit	subreddit_num_posts	Total number of posts in subreddit history
subreddit	subreddit_min_date	Day, Month, Year of first post in subreddit
subreddit	subreddit_max_date	Day, Month, Year of latest post in subreddit

7.2.1.1 Author feature set

Author features were generated for AO frequency, volume of content on Reddit, and length of time on Reddit. For each author, their total frequency of each of the four AOs were created as features. The distributions of unique authors is shown below in Figure 7.2, with authors segmented into buckets to increase readability. The numeric values for this plot, and other plots in this chapter without numeric labels, may be found in Appendix A.

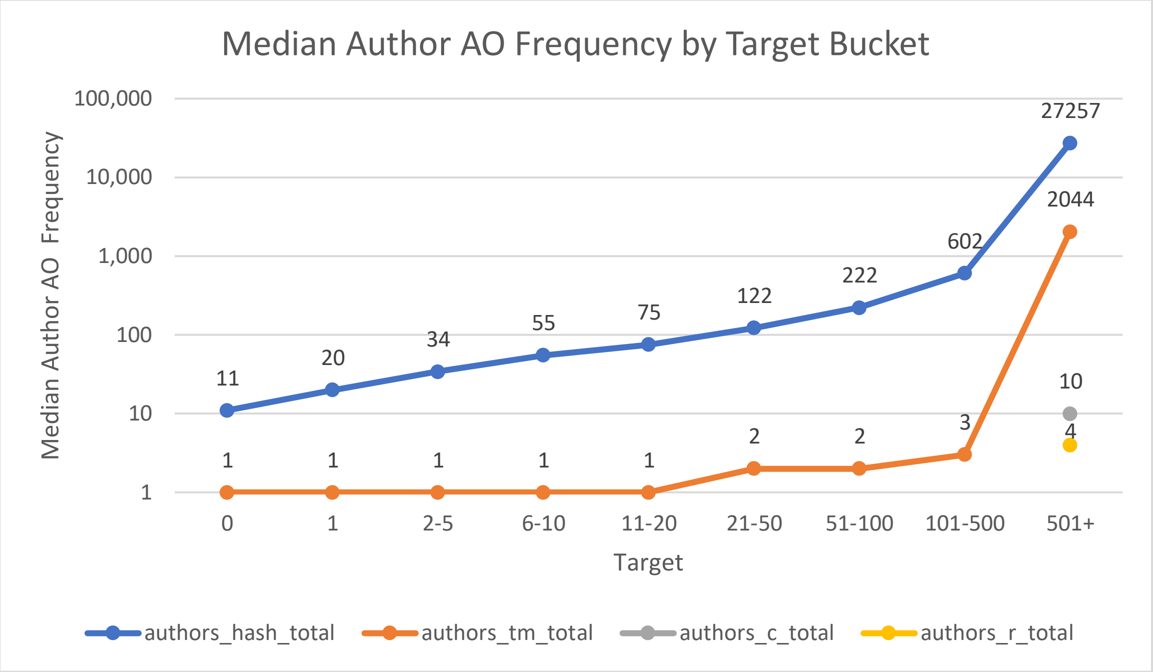
Figure 7.2: Distribution of author AO frequency for regression task



All of the AOs have the highest number of authors with a global AO frequency of zero, except for the trademark, where the highest volume of authors have a global frequency of one.

The median AO frequency features are shown below in Figure 7.3 for each of the target buckets, where the x-axis is the target for a given row in the dataset, and the y-axis is the median feature value for all the rows which fall into that target bucket. For example, for rows which have a target value of 0, the median value for the authors_hash_total feature is 11.

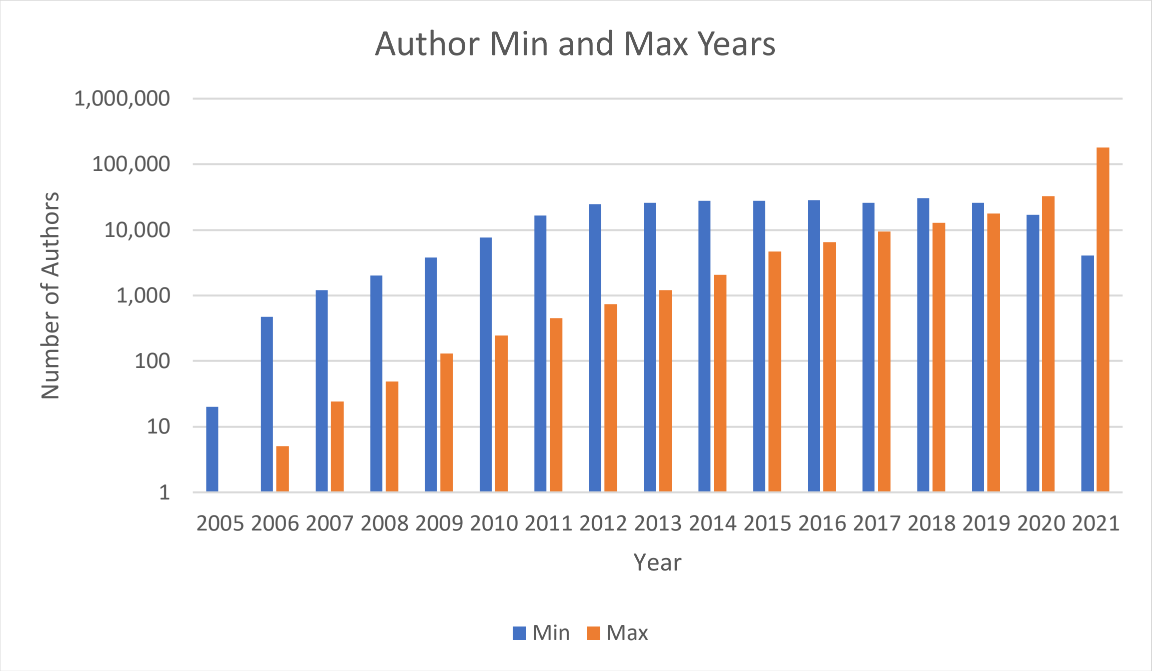
Figure 7.3: Author median AO features by target bucket for regression task



For the hashtag, there is a fairly straightforward relationship between the target and the global hashtag value. As the target increases, so does the total number of hashtags used by a given author. However, for the lower frequency AOs, it is only when the target is extremely high that the median rises above 0. In other words, it is not as simple as saying that the more times an author has posted AOs on Reddit, the more AOs they have posted in a specific subreddit.

Additionally, for each author, the date of their first and last posts on Reddit and the date of their first and last posts in a given subreddit were created as features. The features were split into 3 numeric columns: year, month, and date. The distribution of min and max years on Reddit are in Figure 7.4 below, where the Min columns show the number of authors who had their first post on Reddit in a given year, and the Max columns show the number of authors who had their last post on Reddit in a given year.

Figure 7.4: Author min and max year on Reddit for regression task

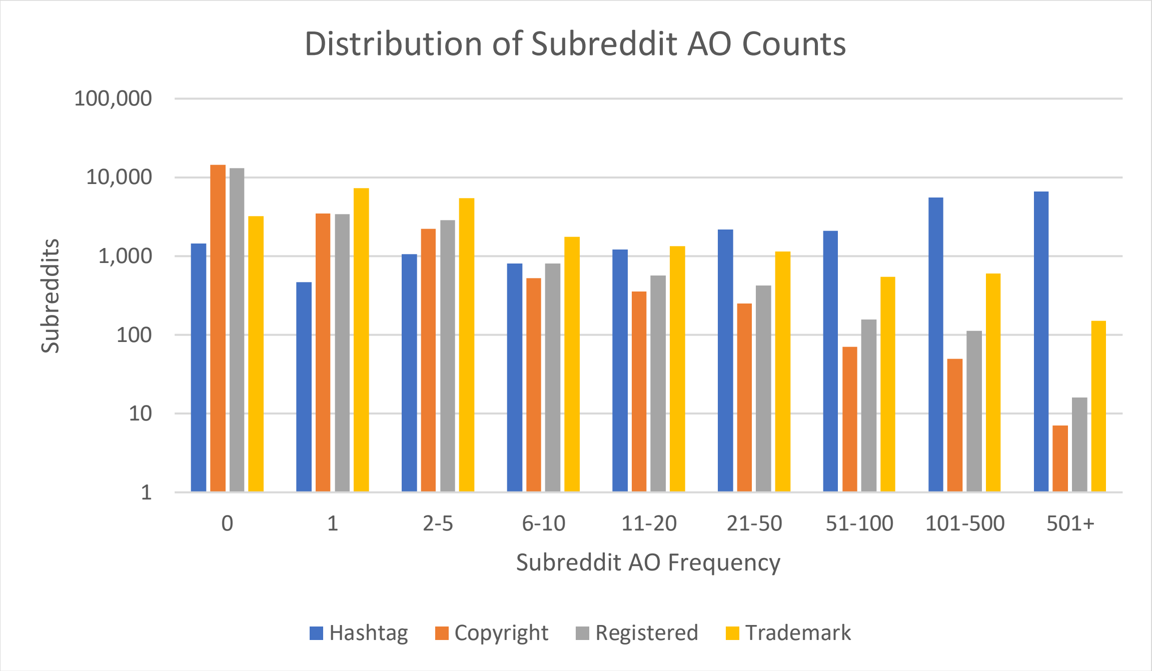


The potential insight related to this feature has to do with the length of time a user has spent on Reddit as well as within a given community. If AO usage is a community norm, or even a Reddit norm, then it is possible that exposure or length of time in a given community would be an important predictor of AO use. Additionally, by using the dates of both first and last posts, the model may take into account the specific timespan someone was on Reddit.

7.2.1.2 Subreddit feature set

Subreddit features were also generated for AO frequency, volume of content on Reddit, and length of time on Reddit. As with authors, the four total AO frequencies for each subreddit were created as a feature. The distribution of this feature is below in Figure 7.5.

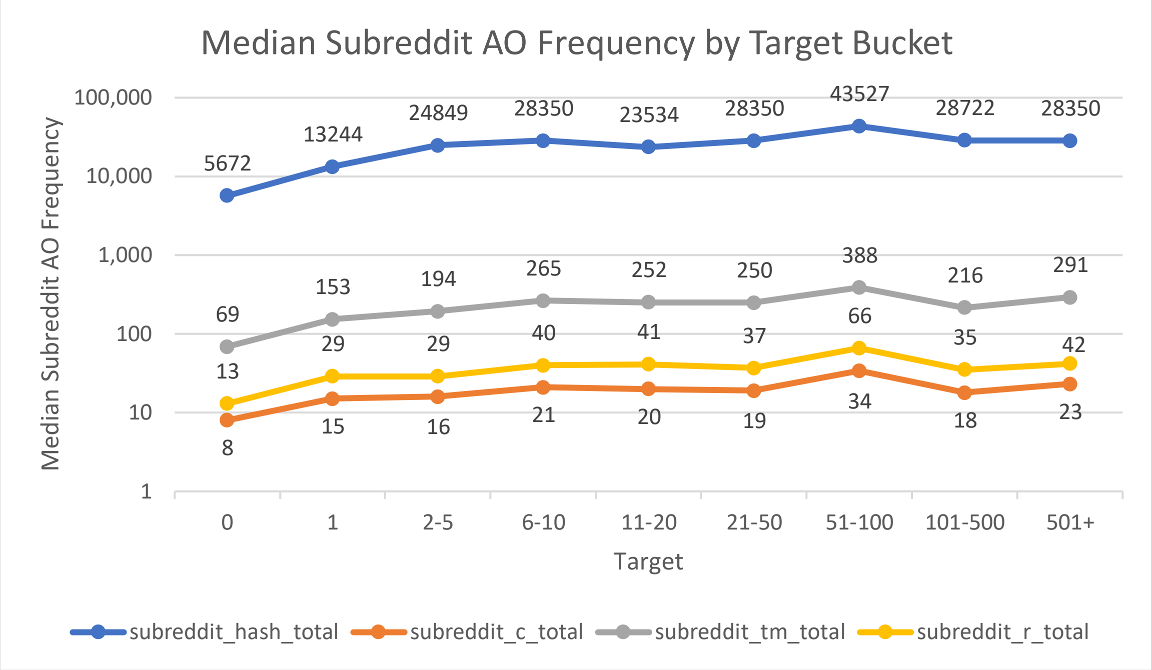
Figure 7.5: Distribution of subreddit AO frequency for regression task



Notably, the hashtag has the largest volume of subreddits in the 501+ bin, which is in contrast to the other operators, which show a steady decrease in subreddit count as the frequency in the bins increases. The dataset had almost 10,000 subreddits containing more than 500 hashtags, but less than ten containing more than 500 copyright symbols.

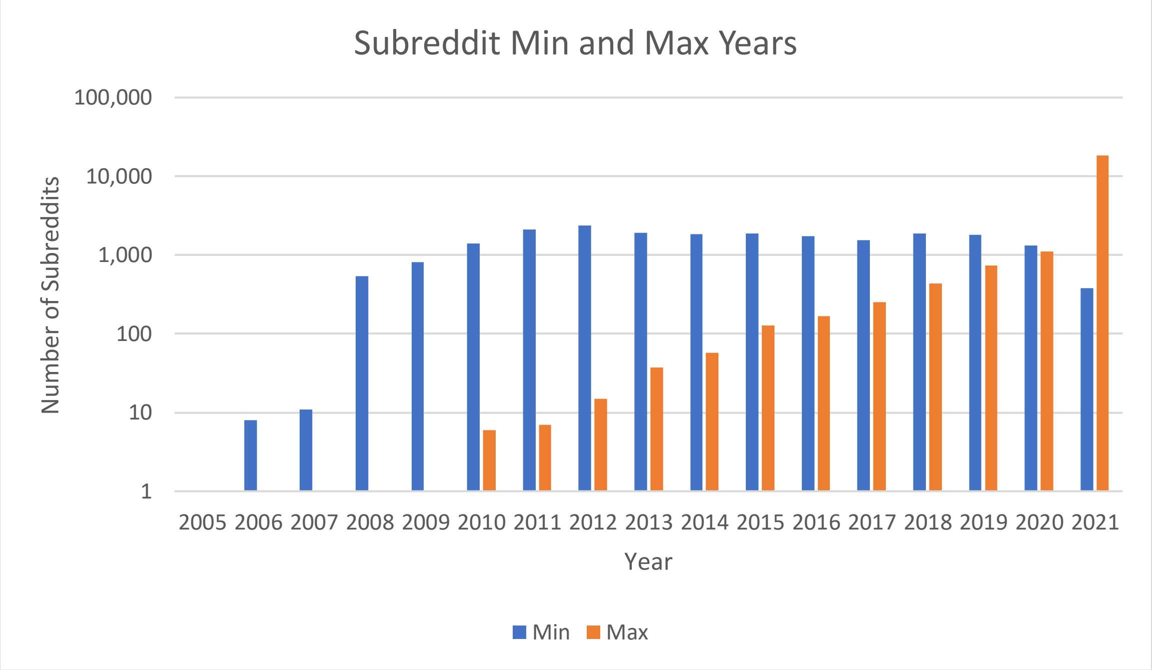
The subreddits also behave differently in terms of medians by target, demonstrated in Figure 7.6 below. While there does appear to be a relationship between the target and global AO counts, it is much more gradual than the Author counts. This is likely due to the fact that subreddits, as aggregates of author comments, contain a higher volume of comments.

Figure 7.6: Subreddit median AO features by target bucket for regression task



Finally, as with author, the first and last date comments were ever posted in the subreddit were turned into features. The distribution of this feature, by min and max year, is below in Figure 7.7.

Figure 7.7: Subreddit min and max years on Reddit for regression task



While some subreddits became inactive as early as 2010, the majority of subreddits in the dataset were still active through 2021, the final year in the dataset. It should be noted that 2021 only contained 6 months of data, which likely accounts for the decrease in authors who made their first comment in 2021 compared to 2020. Subreddit min and max dates were seen as valuable features, because subreddits which have existed longer, or during specific time periods, may have been more likely to develop community-based norms and a shared repertoire.

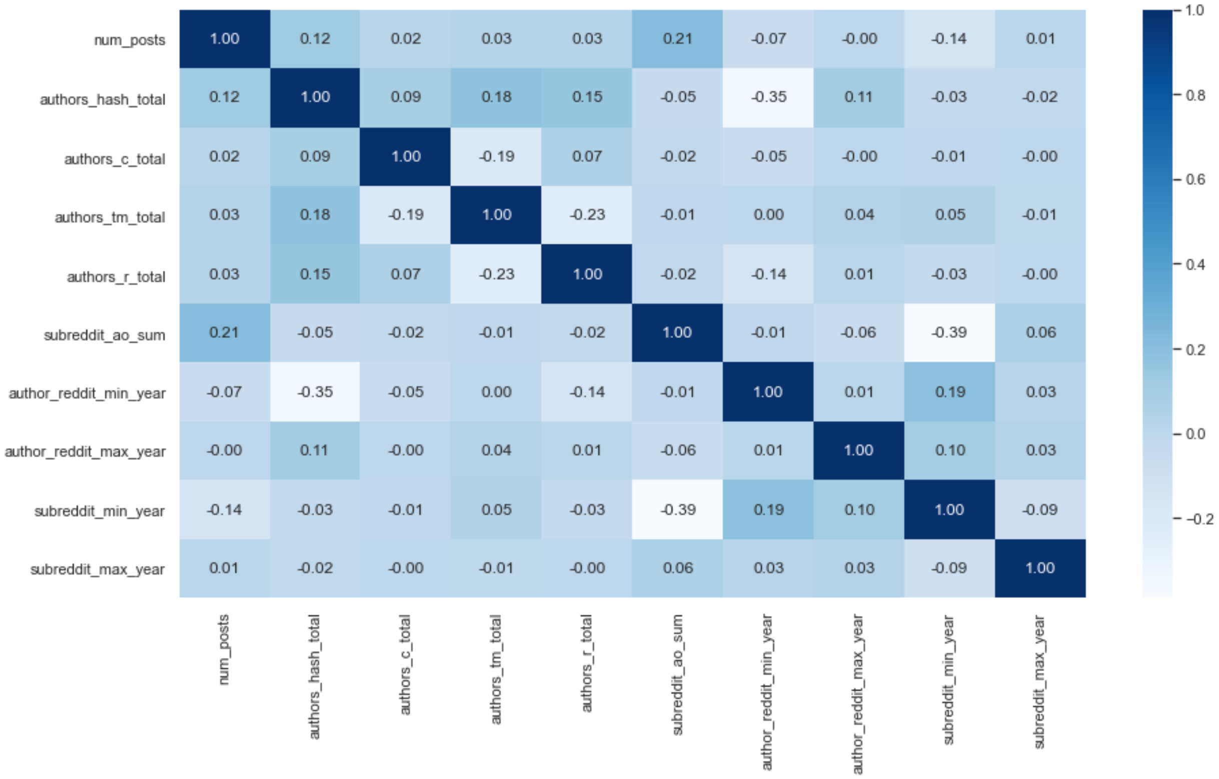
7.2.2 Feature selection

There were two modeling phases. The first phase iterated through different combinations of feature sets but did not generate feature importance ratings. The second phase was focused on generating feature importance rankings. In this phase, correlation between features was an important concern, since highly correlated features can lead to issues with both linear regressions and random forests. Therefore, the first step in this analysis was to remove any features that were highly correlated with others.

To support this second phase of modeling, a correlation matrix was generated across all the features. Many of the features were not normally distributed. As such, Spearman's correlation was used, as this is better for data that is not normally distributed (De Winter et al., 2016). After the initial feature generation there were 26 features (Table 1). The *total* features (e.g., *authors_hash_total*), which tracked the total unique occurrences of an AO, and the *comments* features (e.g., *authors_hash_comments*) which tracked the total unique comments containing an AO, were highly correlated (>0.9) for both authors and subreddits. As such, the eight *comments* features were dropped. The global *number of posts* features and *word count* features were both highly correlated (>0.7) with the *authors_hash_total* feature, and as such the former two were dropped. The *author_subreddit_min_year* feature was correlated with the *author_reddit_min_year* (>0.5) feature, so the former was dropped. The subreddit AO *total* features were all highly correlated with each other (>0.9), so I created an aggregated *subreddit_ao_sum* to sum the values of all 4 AO *total* features. Finally, this aggregated *subreddit_ao_sum* feature was strongly correlated with *subreddit word count*. As AO behavior was seen as more important to the model goals, *subreddit_word_count* was removed. In addition to removing highly correlated features, the *day* (date of month) and *month* features were removed, as they were not seen as sufficiently interpretable.

The correlation matrix for the final 10 features is below in Figure 7.8. The maximum absolute correlation value among the features was 0.39.

Figure 7.8: Correlation matrix for remaining features in the regression models



	num_posts	authors_hash_total	authors_c_total	authors_tm_total	authors_r_total	subreddit_ao_sum	author_reddit_min_year	author_reddit_max_year	subreddit_min_year	subreddit_max_year
num_posts	1.00	0.12	0.02	0.03	0.03	0.21	-0.07	-0.00	-0.14	0.01
authors_hash_total	0.12	1.00	0.09	0.18	0.15	-0.05	-0.35	0.11	-0.03	-0.02
authors_c_total	0.02	0.09	1.00	-0.19	0.07	-0.02	-0.05	-0.00	-0.01	-0.00
authors_tm_total	0.03	0.18	-0.19	1.00	-0.23	-0.01	0.00	0.04	0.05	-0.01
authors_r_total	0.03	0.15	0.07	-0.23	1.00	-0.02	-0.14	0.01	-0.03	-0.00
subreddit_ao_sum	0.21	-0.05	-0.02	-0.01	-0.02	1.00	-0.01	-0.06	-0.39	0.06
author_reddit_min_year	-0.07	-0.35	-0.05	0.00	-0.14	-0.01	1.00	0.01	0.19	0.03
author_reddit_max_year	-0.00	0.11	-0.00	0.04	0.01	-0.06	0.01	1.00	0.10	0.03
subreddit_min_year	-0.14	-0.03	-0.01	0.05	-0.03	-0.39	0.19	0.10	1.00	-0.09
subreddit_max_year	0.01	-0.02	-0.00	-0.01	-0.00	0.06	0.03	0.03	-0.09	1.00

7.2.3 Results

As previously stated, I chose linear regression and random forest regression for this task due to their high interpretability. For the linear regression, features were scaled to have a mean of zero and standard deviation of 1. This enables more direct comparisons of the relative impacts of each feature's coefficients. After scaling, the data was segmented using random sampling with 70% of the data sampled for training and the remaining 30% used for testing. The model was trained on the training sample, then used to make predictions on the testing sample. All reported results are with respect to the predictions made on the test data.

In the first phase of modeling, I iterated through combinations of feature sets in order to explore how the author, subreddit and cross (author x subreddit) feature sets would impact the model. All the original features were used, with none removed due to correlation. The results of these iterations are in Table 7.2 below for both linear regression and random forest. Root mean squared error (RMSE) was chosen as an evaluation metric because it gives a higher weight to large errors. In this sense, it highlights how large the disparity between the prediction of the model and the actual target was. Since for this model, a prediction of 3 AOs when the target is four AOs would still be considered a fairly good prediction, RMSE was seen as appropriate.

Table 7.2: Feature set results for regression task

Feature Set	**Linear Regression RMSE**	**Random Forest RMSE**
Author Set	12.07	12.11
Subreddit Set	11.50	11.53
Cross Set	9.44	51.19
Author + Subreddit	12.09	88.05
Author + Cross	10.20	14.13
Subreddit + Cross	9.46	47.94
Author + Subreddit + Cross	10.22	20.31

The linear model performed best with the cross set, suggesting that those features related to the author's behavior within a given specific subreddit led to the best performance. The author set appeared to negatively impact performance, with the author + subreddit and author + cross combinations both yielding worse RMSE values than the subreddit and cross sets alone. Based on these initial results, it would appear that the cross feature set is the most valuable in terms of predicting number of AOs in a given subreddit, which is perhaps unsurprising given the structure of the dataset. However, looking at only the author and subreddit sets, the subreddit set appears to be the more beneficial of the two. The random forest models performed very differently, and with much higher RMSE values. The performance for the author and subreddit sets for random forest were similar to that of the linear regression. However, for the cross set, Random forest performed substantially worse. Also oddly, the author and subreddit combined features had the worst RMSE at 88.05. Essentially, it appears that for the Random Forest algorithm, none of the combinations of feature sets were very successful, suggesting that perhaps some of the features created noise that prevented the model from performing well. Feature importance rankings were not generated for any of these models, because in this phase I did not control for correlation between the features. As such, these models provided an overview of which features might be useful to the model but did not enable the granularity of comparing specific features.

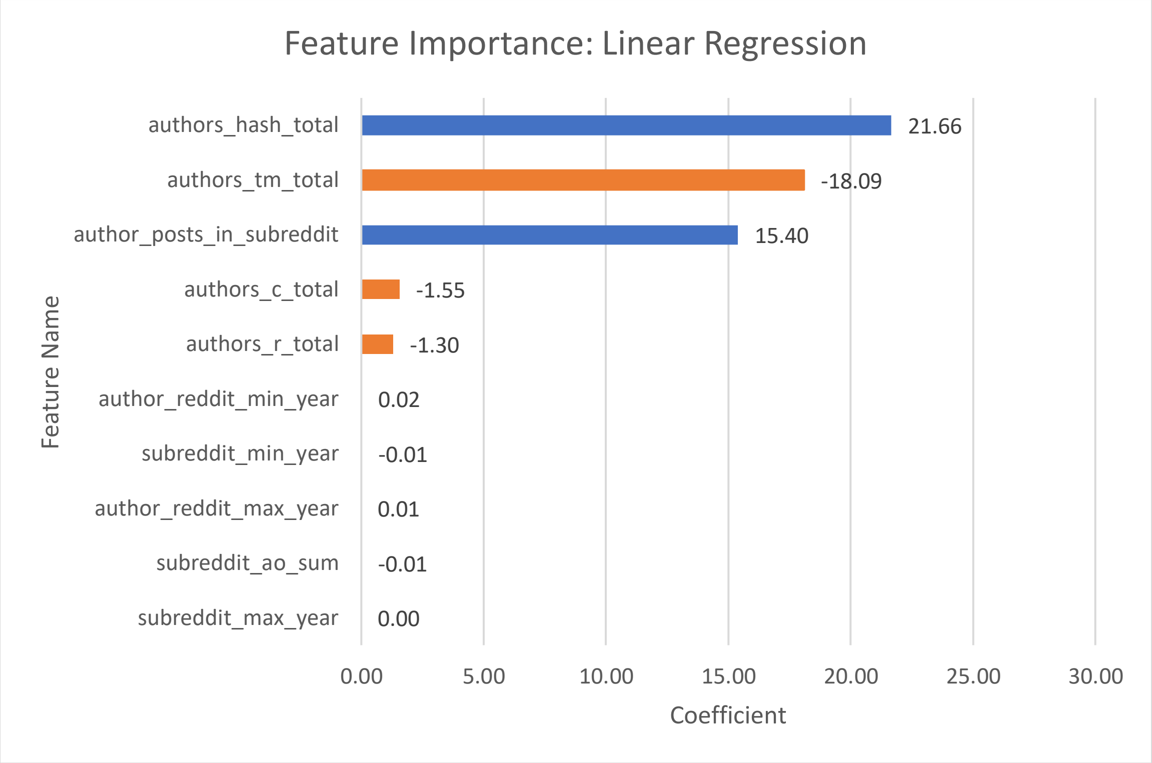
The second phase of modeling allowed me to differentiate between specific features and test combinations from across the various feature sets. These models were trained after removing the correlated features according to the description in the previous section. The performance of the models using all 10 features are below in Table 7.3.

Table 7.3: Regression modeling results with filtered features

Model	**RMSE**
Linear Regression	10.17
Random Forest	25.59

Neither model performed as well as the best of the feature set models, perhaps due to having fewer features. Interestingly, the linear regression model again performed substantially better than the random forest regression model. While neither model performed well, performance of the model was seen as secondary to extracting insights from the feature importance. Therefore, the linear regression model, as the better of the two, was used to rank the 10 features according to their model coefficients. In the graph, negative coefficients are orange, while positive coefficients are blue.

Figure 7.9: Feature importance rankings from linear regression



In the final model, the top two most influential features were related to the two highest frequency AOs: the hashtag and the trademark symbol. The largest coefficient was for authors_hash_total, which had a positive coefficient of 21.66. This would suggest that the model performed best under the assumption that the more hashtags a person has posted on Reddit, the more AOs they have used in a given subreddit. This relationship seems fairly straightforward. By contrast, *authors_tm_total* has a negative coefficient of 18.09, suggesting that the more trademarks an author has used on Reddit, the fewer AOs they may have used in a given subreddit. While this initially seems counterintuitive, it could be an indicator that authors who have a high volume of trademark usage may do so within a small portion of their overall communities. The model may be learning that while high hashtag frequency across the corpus generally relates to higher hashtag use within individual communities, for the trademark this is not always the case. One possible reading of this is that hashtag use is more universal while trademark use is more community specific. The next important feature was the *author_posts_in_subreddit* feature, which has a large, positive coefficient. If an author was more engaged with a particular community, they were also more likely to have posted more AOs.

Looking at the next two AO features, the registered trademark and copyright both have small, negative coefficients, suggesting a similar relationship to the target as the trademark, but on a smaller scale, likely due to their low frequency. The date-based features, and the subreddit AO sum, all had extremely small coefficients. Interestingly, although the author feature set was seen to negatively impact the performance in the phase 1 modeling, when looking at feature coefficients, author-related features dominated the top spots. This might suggest that the author features have a large impact on the model but are less consistent in their relationship to the target. The subreddit features may also not have been as useful in aggregate as all four AOs summed together compared to broken out by AO as they were during the feature set iterations. However, it is clear that AO volume and post volume were more effective for model performance compared to the features surrounding tenure on Reddit.

7.3 Classification task

The second task involved predicting whether a given user has ever posted a comment on Reddit that contained an AO. This was conceived as a binary classification task, with the target either being a 1 (yes, the author has posted a comment with an AO) or 0 (no, the author has never posted a comment with an AO). The source dataset was the same aggregated Reddit dataset described in section 7.1. However, rather than comparing author and subreddit features, the main goal of the classification task was to understand the degree to which specific subreddit communities might influence AO usage.

7.3.1 Feature generation

Despite using the same underlying dataset as the regression task, it was necessary to generate different types of features due to differences in the goals of the tasks. In the regression task, an author's AO frequency was encoded as four different features. This was obviously not possible in a task predicting whether or not an author has ever used an AO. Additionally, in the regression task, each row was an author-subreddit combination, such that features specific to a given subreddit were able to be encoded for each row. However, in the classification task, the goal was to make direct comparisons between specific subreddits, rather than subreddit features and author features more broadly. As such, subreddit information was encoded differently for the classification task, such that each row is a unique author and specific subreddits were used as features. Three author-related features were generated: *author_word_count, reddit_start,* and *reddit_end*. These features are described in Table 7.4 below. The Reddit start and end features were encoded differently in this experiment, with the value being number of days after the earliest date of the corpus: 12/12/2005. This was done to allow for more granularity in terms of a start date, rather than using an entire year, as was done in the phase 2 modeling for the regression task.

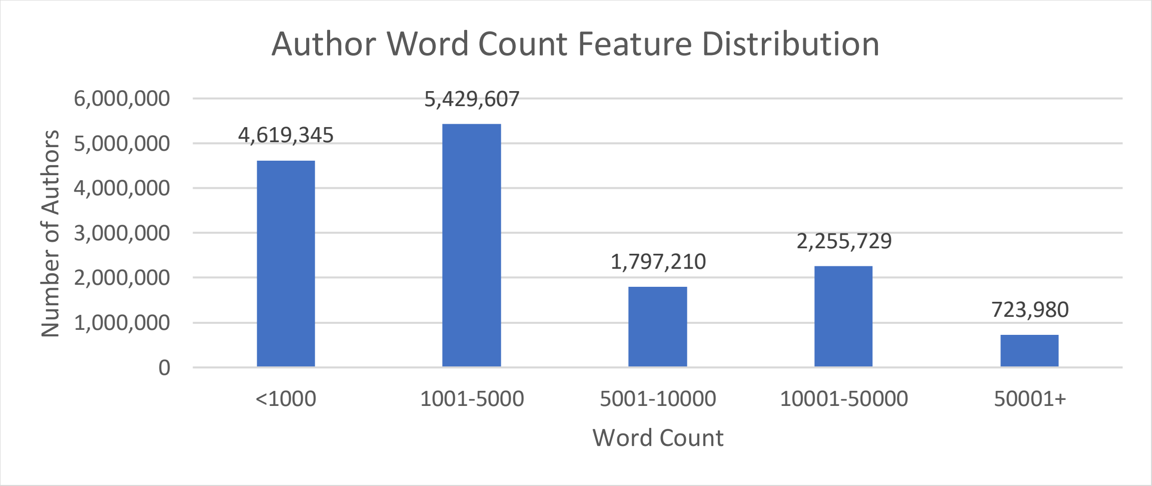
Table 7.4: Classification task features

Feature	**Description**
author_word_count	Total number of tokens in author history
reddit_start	Number of days after first Reddit post when author first posted
reddit_end	Number of days after first Reddit post when author last posted
Subreddit_n*	Number of posts in subreddit

In addition to the three author-related features, subreddits themselves were used as separate features. 1000 subreddits with a high relative frequency of AOs were selected to become features. These subreddits were selected based on two criteria. The first was that the subreddit had at least 1000 comments in its entire history. This was to ensure that spam-filled or brand-new subreddits did not appear in the dataset. 1000 posts was deemed a sufficient threshold for some sort of sustained interest or community. For all subreddits with at least 1000 posts, a weighted frequency was calculated using the less frequent AOs (©, ®, ™). The weighted frequency was calculated as the sum of all comments containing non-hash AOs over the total number of posts in the subreddit. After generating this weighted frequency, the top 1000 subreddits were chosen. This yielded 1000 distinct features, each representing a particular subreddit, with the value being the number of posts an author has made in each community.

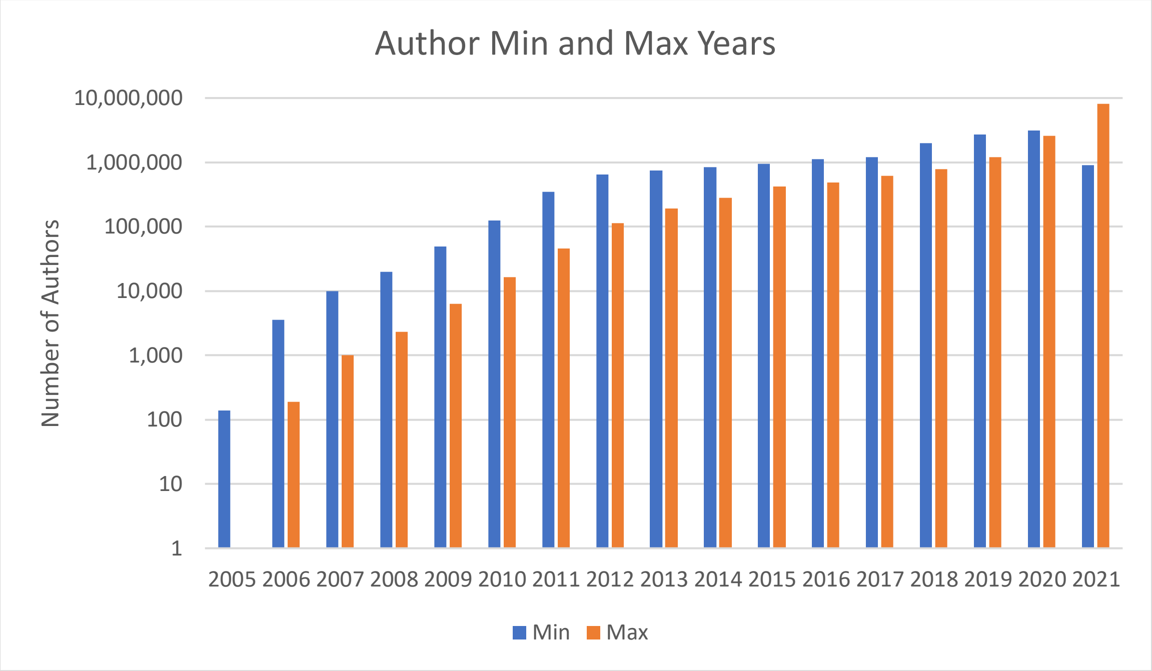
The distribution of the author word count feature is below in Figure 7.10 (non-logarithmic axis). The distribution skews left, with the majority of users having a word count below 5000 words. However, there are many users who have more than 10,000, or even more than 50,000 words in the dataset.

Figure 7.10: Author word count feature distribution for classification task



The author’s first and last dates on Reddit were also used as features. The distribution of these is below in Figure 7.11. As with the min and max features of the regression task, the figure shows the continued growth of Reddit in terms of number of users. Again, it should be noted that 2021 only represents 6 months, rather than entire year.

Figure 7.11: Author min and max years for classification task



As with the regression task, interpretability was a priority. As such, logistic regression and random forest classification were selected to obtain feature importance rankings. For the logistic regression model, the data was scaled to ensure comparable coefficients. Due to the way subreddit features were encoded (sparse vectors), the dataset was stored in a sparse matrix format. Scaling features to a mean of zero and standard deviation of 1 would destroy the sparsity, and potentially cause out-of-memory errors, so all variables were instead scaled to be between 0 and 1.

The data was somewhat imbalanced, with 3,096,186 authors (69.6%) who have never used an AO and 1,351,576 (30.4%) authors who have used an AO, yielding a total of 4,447,762 observations – one record for each unique author in the dataset. The data was segmented using random sampling with 70% of the data sampled for training and the remaining 30% used for testing. As with the regression task, the model was trained on the training sample, then used to make predictions on the testing sample. All reported results in the following section are with respect to the predictions made on the test data.

7.3.2 Results

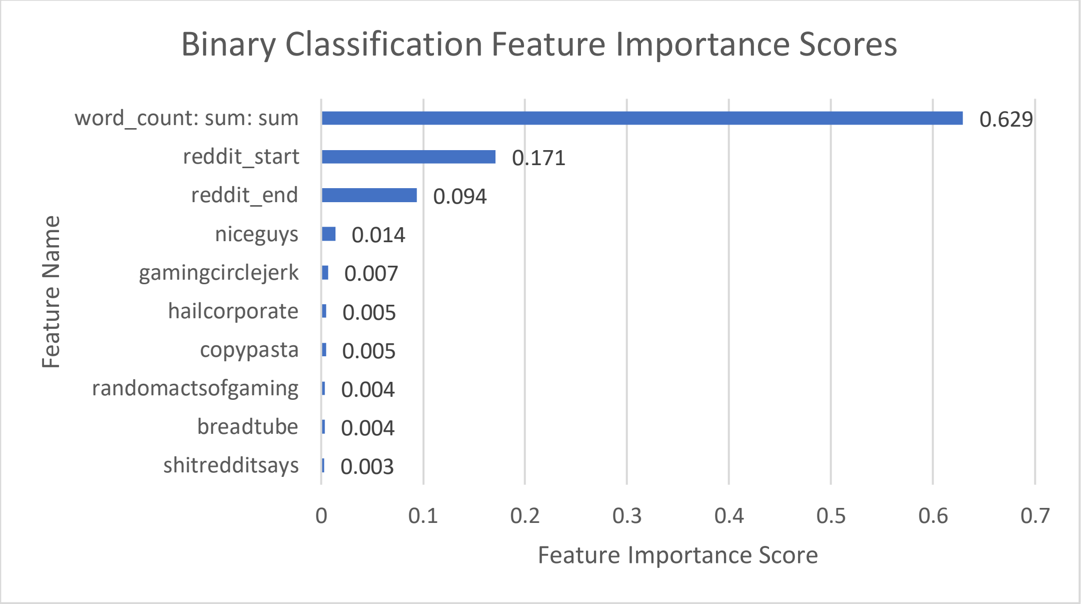
A series of binary classification models were trained on the dataset. The results of the models are below in Table 7.5. When the model makes a prediction, it can fall into one of four categories: true positive, false positive, true negative, and false negative. True positives are instances where the model correctly predicts that an author has used an AO. False positives are instances where the model predicts that an author has used an AO, but they did not. True negatives are instances where the model correctly predicts that an author has never used an AO. False negatives are instances where the model predicts that an author has never used an AO, but they have. Evaluation metrics used were precision, recall, AUC, and accuracy. Precision is calculated as the number of true positives divided by true positives plus false positives. Recall is calculated as the number of true positives divided by true positives plus false negatives. AUC is the area under the receiver operating curve (ROC). An AUC of 0.5 is roughly equivalent to a random guess. The higher the AUC, the better the model is at distinguishing between the classes. Accuracy is calculated as the number of correct predictions divided by the total number of predictions. Accuracy alone was not considered a very useful metric due to the imbalanced data.

Table 7.5: Classification modeling results

Model	**Precision**	**Recall**	**AUC**	**Accuracy**
Logistic Regression	0.66	0.32	0.63	0.74
Random Forest	0.74	0.53	0.73	0.80

While none of the models performed well, they all performed better than chance with AUC scores greater than 0.5, suggesting that there is some relationship between the features and the target. The best model was the Random Forest model. The next step was to produce the feature importance rankings. In contrast with a linear or logistic regression approach, tree-based models do not produce coefficients. Instead, feature importances are calculated by the normalized total reduction in the ensemble's loss function that results from splitting on each feature. Importantly, due to the nature of random forests, these feature importances cannot indicate directionality, as with linear regressions. A high importance for one feature does not indicate anything about the exact relationship between that feature and the target. It only indicates that this feature is useful for correct classification. The feature importance rankings for this model are below in Figure 7.12.

Figure 7.12: Feature importance rankings from random forest classifier



The author word count was the most important feature in predicting whether or not someone has used an AO. The next two features were reddit_start and reddit_end, suggesting that an author’s tenure on Reddit was beneficial in predicting whether or not the person has ever used an AO.

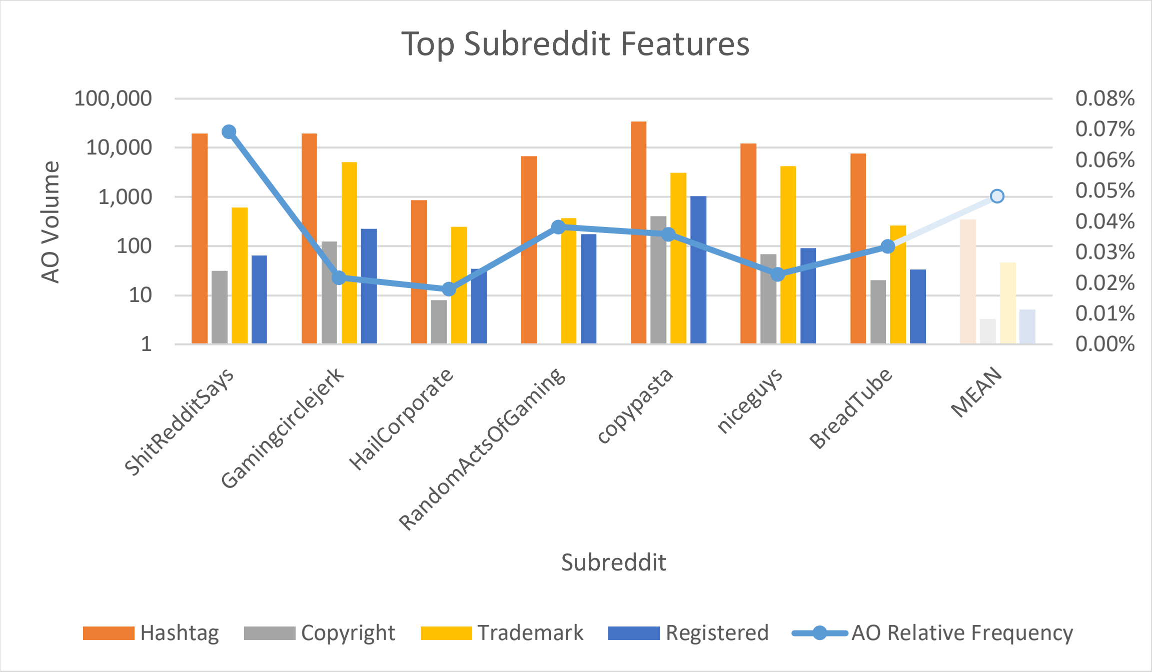
There is a substantial drop off in score before the first subreddit feature in the feature importance, suggesting that in general, the subreddit features were not as helpful. However, it is interesting to note which subreddits were ranked highest in terms of their contribution to the model. The top 7 subreddits were niceguys, Gamingcirclejerk, Hailcorporate, copypasta, RandomActsOfGaming, BreadTube and ShitRedditSays. A description of the thematic content of each subreddit is below in Table 7.6.

Table 7.6: Thematic content of important subreddit features from classification task

Subreddit	**Description**
niceguys	Stories and discussions of encounters with nice guys (e.g., men who think kindness should be repaid with romantic or sexual affection)
Gamingcirclejerk	A meme community for complaints and parodies of games and gaming culture
HailCorporate	Discussions of unwitting and legitimate advertisements for brands
copypasta	Stories that frequently get copied and pasted around the internet
RandomActsOfGaming	A subreddit which hosts giveaways for games and game products
BreadTube	A place for posting content that is counter to politically mainstream internet culture
ShitRedditSays	A place for mocking comments and posts made by other Reddit users

The AO volumes of the seven subreddits are shown in Figure 7.13 below, as well the mean for the 1000 subreddit features on the far right. The relative frequency is also displayed in the graph on a secondary (non-logarithmic) axis.

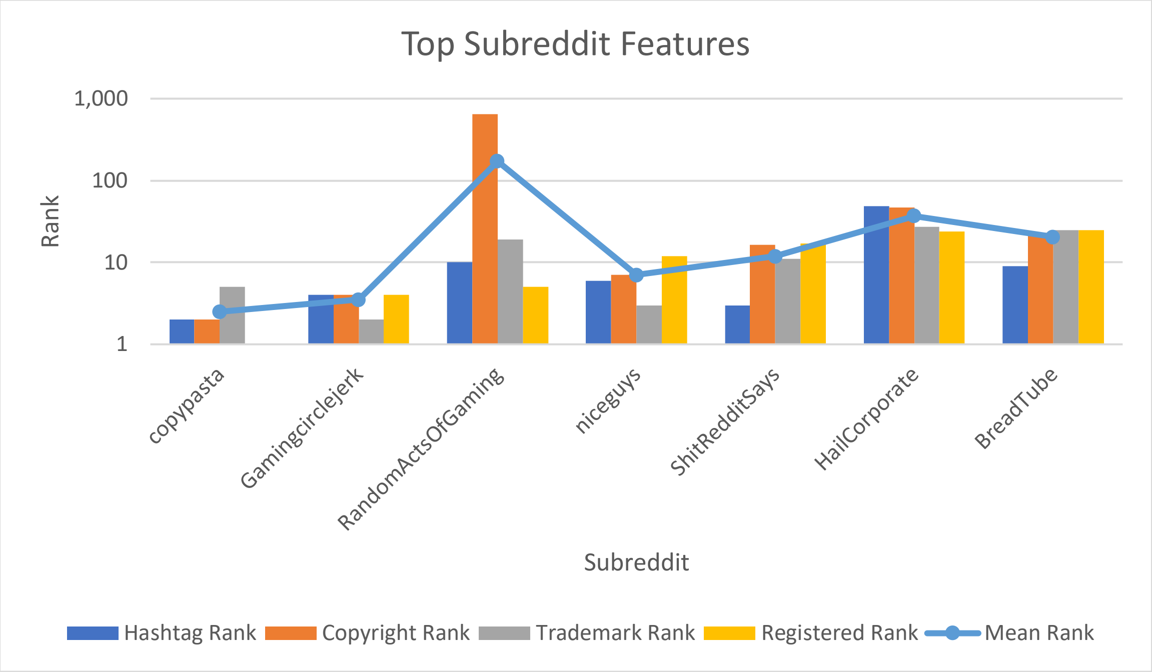
Figure 7.13: AO frequency for top subreddit features from random forest classifier



The relative AO frequency for each of the top subreddits is below the mean, except in the case of *ShitRedditSays*, however the raw counts for each of the AOs is above the mean in all cases. Although AO-related features were not passed to the model, it stands to reason that AO frequency is what made these subreddits more valuable.

In Figure 7.14 below, the ranking among the 1000 subreddits by AO frequency is given for the top 7 most important subreddits. The lower the ranking, the higher the frequency (e.g., a ranking of 1 corresponds with having the highest raw frequency of that AO out of all subreddits).

Figure 7.14: AO ranking for top subreddit features from random forest classifier



Each of the six subreddits in the top feature importance rankings was in the top 100 subreddits for every AO, with the exception of RandomActsOfGaming and the copyright symbol, which was ranked 652/1000. The most predictive subreddits, in other words, were communities that had a high comparative frequency across all 4 of the operators.

7.4 Conclusion

The goal of the two machine learning tasks was to identify features that might enhance our understanding of what drives AO usage. In the regression task, I complete two phases of training. In the first phase, which used feature sets, it appeared that the subreddit features were more important than the author features. The model trained on only subreddit features had a lower RMSE compared to the model with subreddit + author or just author. However, in the second phase, where I generated feature importance values, the author features were much more important to the model compared to the subreddit features. It is possible that the author feature set as a whole performed badly due to some noise generated by the broader group of features. In other words, to the question of whether a subreddit community or an author's individual behavior contributes more to the model performance, going off of the feature importance, author related features were more predictive. This makes sense since even in communities with a culture of AO use, it is often a small subset of users who actually employ them. However, it is clear that the subreddit features were of some value, since the subreddit feature set model performed at a similar RMSE. Among the four operators, there was a clear relationship between frequency and feature importance, with the hashtag being the most important, followed by the trademark. However, the behavior of the registered trademark and copyright symbol were more interesting. Their frequencies in the corpus are quite similar, but the copyright symbol was less influential to the model, and additionally had a negative coefficient, suggesting that being a copyright user was not a very powerful indicator of broader AO use. Given what I saw in the corpus analysis, where copyright symbols were sometimes used in more traditional ways, this makes sense.

The key insight from the classification task was that individual author behavior was substantially more important to the model than the subreddit features. However, for the subreddit features that were important, almost all of them demonstrated a high frequency across all 4 AO types. Because the target variable relates to the author, it makes sense that author features were more predictive across both tasks compared to subreddit features. However, it is interesting to note that the model performance picked out subreddits with high relative frequency and a culture of use across all four AOs as being most useful.

CHAPTER 8

CONCLUSIONS

In this dissertation, I have provided a descriptive account of four artificial operators (AOs) on Reddit. In chapter 3, I presented a quantitative corpus analysis of AO use on Reddit alongside a case study of three subreddits with high AO frequency. In chapter 4, I argued for the framing of AOs as stance markers. In chapter 5, I showed that AOs interact with discourse-oriented adjectives and adverbs via scalar modification and scalar inversion. In chapter 6, I presented the results of two different surveys conducted with social media users. Lastly, in chapter 7, I presented the results of two machine learning models which aimed to predict AO use. In this chapter, I summarize the findings of the dissertation. In section 8.1, I explicitly relate the findings to the four research questions I posited at the beginning. I also place the findings from the various chapters in dialogue with one another, to produce a cohesive analysis of AOs on Reddit. In section 8.2, I discuss the implications of this work and how it relates back to current theories of Computer-Mediated Communication (CMC). Finally, I conclude the dissertation in section 8.3 by acknowledging some limitations of the project and making recommendations for areas of future work.

8.1 Findings

In this dissertation, I studied AOs using a mixed-methods approach. First, I used corpus analysis to analyze a corpus of Reddit comments spanning from 2005-2019. This allowed me to perform a diachronic study of AO use on Reddit. I performed frequency analysis to see changes in raw and relative frequency of AOs. I examined affixed unit length over time and found that affixed units are short, but that they have increased in length slightly over time. I also used automatic POS tagging to observe changes over time in the distribution of various POS categories. I found that AOs largely affix to NPs but are capable of affixing to almost any POS category. Across this analysis, I found that the hashtag behaved differently from the other operators. In addition to examining broad trends in the corpus, I performed a closer analysis of 3 subreddits where AO-use is relatively high. I examined *transgendercirclejerk* (a meme/parody community for trans people), *2b2bt* (a community for a Minecraft server), and *COMPLETEANARCHY* (a meme community for political anarchists). Performing direct comparisons between these communities revealed differences in how each community employed AOs. I argued that this aspect of AOs makes them very compelling evidence for identifying Communities of Practice (Lave and Wenger, 1991; Wenger 1998), which have a shared repertoire.

In the qualitative analysis, I posited that AOs may be best captured by a typology rather than discrete categories. I created a typology along the lines of +/– evaluative and +/– distanced, showing that these two spectrums effectively capture the impact of the operator on the affixed text. I used the Du Bois (2007) stance triangle to demonstrate how AOs are used as stance markers. I showed that AOs play a role in evaluating a stance object, positioning an author in relation to a stance object, and signaling alignment with a particular group or perspective. Crucially, I adapted the stance triangle such that I reframed positioning as related to critical distancing. I demonstrated the ways that authors use AOs to enact various stances. However, I did not argue that stance marking was a comprehensive approach to the effects that AOs have, as they are clearly capable of performing a variety of functions at once.

One of these additional functions is related to scalar modification, when the AO is affixed to adjectives and adverbs. I discuss discourse-oriented adjectives, which invoke an implicit scale and can only be interpreted via contextual information. I argued that when AOs and discourse-oriented adjectives are combined they create an upscaling effect that leverages the context-based meaning to create a more emphatic effect. I also showed that sometimes this upscaling effect can be inverted when combined with critical distancing.

In my previous work, a major concern in performing analysis on AOs has been the inherently subjective nature of qualitative analysis. In combination with the few existing studies on AOs, this has led to questions of undue researcher bias in interpreting the qualitative results. As such, a goal of this dissertation was to receive input from external participants in order to understand how users of social media might perceive these operators. Two surveys were designed and administered. The first survey was designed to address the hypothesis that there is a perceptible difference between genuine and figurative AOs. Results showed that there was a statistical difference between the control and experimental items. In other words, participants were able to differentiate between genuine and figurative operators at a statistically significant level. The first study was also leveraged to understand what functions participants would attribute to these operators. The results of the first study yielded six communicative functions that participants had given: sarcasm, quotation, familiarity, humor, emphasis, and quintessential. These six functions were used to aid the design of a second experiment. In addition to understanding how participants perceive figurative AOs, the study also provided insight into how participants perceive genuine AOs. Participants labeled hashtags as being used to create trends or make a post findable, trademark and registered trademark symbols as signaling ownership by a company, and copyright symbols as signaling legal protection.

The goal of the second experiment was to address whether participants would perceive differences between the meanings contributed by AOs under controlled conditions. The second experiment was also to see if the functions yielded by the first experiment would hold up under further scrutiny or show preference for particular operators. The results of the experiment did not show any significant differences between the operators. However, they did show differences between the functions, with significantly higher ratings for attitude over quotation across all four operators. In summary, participants did not perceive differences between the operators, but they did appear to associate AOs more with attitude, sarcasm, and familiarity compared to quotation, quintessence, and humor. These findings aligned with the qualitative analysis, which attributed the functions of AOs to be on a spectrum of evaluation and sarcasm. The familiarity function additionally aligns with the qualitative analysis as I have argued that within CoP the AOs are used to pick out a specific shared referent, and that this a manifestation of shared alignment that occurs during stancetaking.

The last form of analysis used in the study involved using machine learning to predict user behavior around AOs. In this study I performed two machine learning tasks. The first was a regression problem which aimed to predict how many AOs a user has posted in a particular subreddit. For this problem, linear regression performed better than random forest across dozens of iterations. As such, the best linear regression model was used to generate feature importance rankings, which were derived based on the size of the model coefficients. While the model's performance was less than desired, in the final model, author-related features were more important than subreddit-related features. Of particular note, the hashtag was positively correlated with the target while the trademark was negatively correlated with the target. This may suggest that authors use the hashtag more evenly across communities, but use the trademark largely in specific communities. This finding corroborates the case study findings from the corpus analysis, where top AO users from the case study subreddits did not use AOs at the same rate in their other subreddits. The second machine learning task was a binary-classification task. In this task, the goal was to predict whether a user has ever used an AO on Reddit. Findings suggested that the author related features were more important, however, I noted interesting patterns in which subreddits were most predictive out of the 1000 subreddit features. Rather than it simply being a relationship in which subreddits with more AOs are more predictive, certain subreddits with a broader range of use across all four AOs were most important to the model. One interpretation of this is that the presence of all four AOs is a stronger indicator of an influential culture of AOs. In other words, in subreddits where AO use is most integral to the community, people adopt and use multiple AOs.

In summary, in this dissertation I have posited that AOs have shared qualities which motivate their study together as a group. I have argued that while the hashtag is used in a broader range of contexts compared to the other operators, it shares a number of overlapping features and functions with them. I have demonstrated that there is a perceptible difference between the genuine functions of these operators and their figurative or communicate functions. I have also demonstrated that under controlled conditions, these operators may be interchanged without a statistically significant impact on their meaning. I suggest that their function exists on a spectrum of evaluative meaning and critical distancing, but that they are often used within communities to enact various stances or upscale the meaning. I also argue that due to their low frequency, they represent a valuable instance of shared repertoire within online communities to demonstrate the existence of CoP. In the following section, I will tie these findings back to the specific research questions I posited at the beginning of this dissertation.

8.1.1 Research questions

> RQ1: What are the linguistic features that condition the use of artificial operators? How have these changed over time? What drives their use?

The first research question was addressed by the corpus analysis. Operators were found to be extremely flexible, but with a preference for relatively short NPs. Despite this preference, they were found in a broad variety of contexts with seemingly no constraints on how they are able to be used. Other than the hashtag, AOs have not seen much of a change in terms of relative frequency on Reddit. However, they have undergone a slight broadening in terms of the POS they are affixed to. In the corpus analysis, I argue that membership in a CoP is largely what drives their use.

RQ2: What are the pragmatic functions of artificial operators as used on Reddit?

The second research question was addressed by the qualitative analysis. By looking at select examples from the corpus, performing additional analysis on particular affixes in particular communities, I argued that AOs function as stance markers. I also demonstrated that they perform scalar modification and inversion when interacting with adjectives and adverbs. I do not claim that these two functions encompass everything that AOs do, but instead argue that these two functions successfully explain part of what they do.

> RQ3: What are the differences in meaning contributed by ‘genuine’ vs. ‘figurative’ artificial operators? What are the differences in meaning contributed by different artificial operators?

The third research question was addressed by the experimental analysis. Survey one demonstrated that participants do perceive a difference between genuine and figurative uses of the operators. Survey two did not yield a significant difference between the operators, but did indicate a significant difference between the functions. In other words, my findings suggest there is a difference between genuine and figurative operators, but do not suggest a difference among the operators themselves.

> RQ4: Can we predict AO use on Reddit with machine learning models? Will author-related features or subreddit-related features be more important to the models?

The last research question was addressed by the machine learning chapter. Using feature importance scores from two different machine learning models, I found that author-related features were more important to the models.

8.2 Implications

This dissertation has a number of important implications. I have proposed AOs as CMC cues and markers of CoP. For the category of AOs, I have successfully demonstrated value in studying these operators together. I have shown that they share metadiscursive and morphemic qualities, and I have proved that they each show extension from a genuine to figurative use. This, alongside the aforementioned survey results showing no significant differences among the operators, suggest that these operators have much in common and benefit strongly from being analyzed as a group. I propose that the morphemic quality and metadiscursive component of these operators is what causes this shared behavior, and that there are likely other operators that share these features beyond the four I have discussed in this dissertation.

This dissertation was motivated by the substantial body of work on CMC cues. AOs activate a great deal of contextual information. During any given use, AOs invoke the genuine meaning, the figurative meaning, the meaning in a specific community, as well as the meaning in a specific discourse context. I have suggested that AOs perform multiple functions simultaneously, such as stance marking and scalar modification. I have further proposed that as a resource, the meaning of AOs exists across spectrums of evaluative meaning and critical distancing. With the written, nonlinguistic origins of AOs, they do not appear to correlate directly with a single, equivalent paralinguistic cue in spoken language. Instead, they provide examples of CMC cues acting as new linguistic resources, which are in some cases being leveraged in spoken language (Scott, 2018).

Examining AOs in conjunction with CoP has demonstrated that CMC cues provide a valuable entry-point for identifying and analyzing CoP. CMC cues have been shown, in this paper and others, to be highly context-dependent and to indicate shared community membership. CMC cues present a very useful avenue for demonstrating the shared repertoire required to constitute CoP.

Lastly, the machine learning approach was beneficial and suggests that other CMC cues may be used in this kind of study to examine potential latent insights. With machine learning as such a powerful tool, it behooves linguists to learn more about it and leverage it in our research when appropriate.

Additionally, the mixed-methods approach of this dissertation yielded synergy between the results and ultimately, a more robust set of results. The corpus analysis in chapter 3 underpinned the surveys designed and described in chapter 6. The surveys helped to corroborate and reinforce the results of the qualitative analysis. The qualitative analysis guided the design of the machine learning experiments, and the machine learning insights complemented and enhanced the findings of the corpus analysis. While adopting so many distinct approaches likely limited the depth of the work within each approach, the breadth accomplished by using so many approaches allowed for a more comprehensive descriptive analysis. Because of the volume of data available, CMC studies in particular will benefit from this kind of multi-method approach.

8.3 Limitations and future work

There were a number of limitations to this dissertation that prevented me from investigating all aspects of the data. The corpus analysis of Reddit data was extremely beneficial due to its comprehensive nature. However, I underestimated the difficulties of working with such a large dataset. Future work on AOs on Reddit should continue to dive into the massive volume of data that this dataset provides, and perhaps do more analysis on top affixed units across the entire corpus. Additionally, future work would benefit from selecting a small group of examples that would enable hand labeling of linguistic components that may not be detectable via automatic methods.

Future work on AOs should expand the scope of the data sources by examining AO use on different social media sites such as Twitter, Tumblr and Facebook. Because AOs are highly contextual and linked to community practices, examining their use in other contexts will enable a deeper understanding of how different communities leverage these operators in different ways. Additionally, much of the scholarship on hashtags and AOs thus far has focused on English language uses. Future work should examine how operators are used cross-linguistically.

The scope should also be expanded by looking into other operators that may be classified as AOs. Two such examples that might benefit from inclusion in the AO category are *mc*- and *.com*. *Mc*- is a prefix commonly associated with the fast-food chain McDonald's, however there are instances on the internet of people applying *mc*- as a prefix in contexts that appear to have nothing to do with the restaurant as in example 211.

211. Damn that ending truly made me **mcsad**

Additionally, *.com* is a suffix that is typically used to indicate a domain name for a website. However, it appears to behave similarly to AOs in being repurposed for a communicative effect as in example 212.

212. Nah, we need this to be **Tested(.com)**

Future work should examine these communicative uses of the *mc*- prefix and *.com* suffix to understand if they behave the same way as the operators in this study. The criteria that I have outlined for AOs is that the operator has origins in nonlinguistic purposes, does not change the syntactic structure of the sentence, and cannot stand alone in the sentence. From a cursory look, it appears that *mc*- and *.com* both meet these criteria. There are likely many others that also meet these criteria.

Lastly, future work should also examine the relationship between other CMC cues and CoP. CMC cues such as initialisms and abbreviations present a potentially powerful area of inquiry, due to the fact that if an initialism or abbreviation is not known or discernable to the reader, a breakdown in communication will occur. Therefore, studying patterns of initialisms and abbreviations in subreddits or other online communities present an excellent opportunity for examining the shared repertoire within CoP.

APPENDIX A: UNDERLYING VALUES FOR FIGURES WITHOUT NUMERIC LABELS

This appendix contains tables with the underlying values for any plots which did not contain numeric labels in the dissertation. The tables are titled with the corresponding figure number and title.

Table A.1: Values for Figure 3.2: AO comments over time (raw frequency)

Year	Hashtag	Copyright	Registered	Trademark
2006	115	9	34	46
2007	442	55	165	172
2008	1126	138	420	739
2009	3676	294	1677	2086
2010	6658	561	2203	4411
2011	23763	1034	3844	7398
2012	64211	1614	6684	14159
2013	136470	2594	11588	25012
2014	276549	3076	15507	38614
2015	405851	13133	21595	60600
2016	526213	8694	32301	86214
2017	574286	9503	28723	133128
2018	873409	8396	30169	141177
2019	1262291	10371	34041	156368

Table A.2: Values for Figure 3.3: AO comments over time (relative frequency)

Year	Hashmark	Copyright	Registered	Trademark
2006	275.6577	21.57321	81.49881	110.2631
2007	255.9076	31.8437	95.53111	99.58395
2008	155.4632	19.05322	57.98805	102.0314
2009	194.8806	15.58621	88.90499	110.5878
2010	137.3093	11.56962	45.43293	90.96898
2011	192.7411	8.386749	31.17859	60.005
2012	246.6754	6.200403	25.6775	54.39374
2013	339.2996	6.449353	28.81075	62.18628
2014	520.02	5.784079	29.15921	72.60938
2015	607.1752	19.64768	32.3073	90.66091
2016	657.8503	10.86889	40.38141	107.7813
2017	590.3069	9.768106	29.52429	136.8419
2018	704.6098	6.773349	24.3384	113.8925
2019	758.7766	6.234119	20.46241	93.99448

Table A.3: Values for Figure 3.4: Unique subreddits with AO comments

Year	Hashmark	Copyright	Registered	Trademark
2006	5	2	3	2
2007	4	3	6	7
2008	62	30	49	51
2009	219	51	105	144
2010	525	106	216	305
2011	1673	215	478	650
2012	3733	442	975	1389
2013	5516	724	1565	2166
2014	8877	974	2183	3225
2015	12170	2125	4124	5100
2016	14596	3370	8960	8021
2017	17273	2313	4208	7057
2018	23588	2214	4604	8442
2019	28070	2910	5455	10790

Table A.4: Values for Figure 3.5: Percentage of subreddits with AO comments

Year	Hashmark	Copyright	Registered	Trademark
2006	0.147059	0.058824	0.088235	0.058824
2007	0.095238	0.071429	0.142857	0.166667
2008	0.023006	0.011132	0.018182	0.018924
2009	0.035563	0.008282	0.017051	0.023384
2010	0.044533	0.008991	0.018322	0.025872
2011	0.063087	0.008107	0.018025	0.024511
2012	0.063862	0.007562	0.01668	0.023762
2013	0.062967	0.008265	0.017865	0.024726
2014	0.070011	0.007682	0.017217	0.025435
2015	0.059187	0.010335	0.020056	0.024803
2016	0.059016	0.013626	0.036228	0.032431
2017	0.068103	0.00912	0.016591	0.027824
2018	0.053809	0.005051	0.010503	0.019258
2019	0.048621	0.005041	0.009449	0.01869

Table A.5: Values for Figure 3.6: Affixed unit length

Year	Mean	Median	Max
2006	1.389831	1	7
2007	1.391975	1	9
2008	1.452739	1	12
2009	1.61562	1	15
2010	1.667897	1	13
2011	1.789247	1	43
2012	1.875444	1	41
2013	1.982102	2	81
2014	2.135485	2	100
2015	2.085356	2	117
2016	2.142006	2	67
2017	2.062155	2	140
2018	1.944485	2	72
2019	1.574323	1	117

Table A.6: Values for Figure 3.7: POS by hashtag

Year	AdjP	AdvP	INTJ	NP	NP+Adv	NP+P	Other	PP	S	VP
2006	4	3	0	121	0	0	7	4	0	36
2007	21	16	8	447	0	1	20	2	6	222
2008	48	29	4	1075	0	5	71	18	6	545
2009	105	61	46	3483	5	4	147	47	314	1531
2010	306	143	94	6894	4	5	392	95	263	2209
2011	1069	639	429	22836	42	34	1374	363	714	6779
2012	3327	2042	1815	63364	206	142	3627	837	1924	14838
2013	9346	5814	4753	138474	390	911	9312	2695	5263	31249
2014	14747	14317	10916	252811	584	2303	17272	7780	11489	55753
2015	20464	19575	9294	409853	915	2245	33506	7012	21011	102499
2016	26279	34436	10217	480462	1532	3159	35409	7875	26303	120157
2017	34652	26298	11968	616020	1744	3386	42862	8050	35542	131247
2018	52832	98935	25106	816033	3621	6279	104016	10761	101446	290502
2019	53910	327788	27682	1009722	2435	9620	371090	12202	71079	764025

Table A.7: Values for Figure 3.8: POS by trademark

Year	AdjP	AdvP	INTJ	NP	NP+Adv	NP+P	Other	PP	S	VP
2006	1	1	1	59	0	0	0	0	0	1
2007	6	1	0	182	0	1	14	1	2	7
2008	28	16	8	772	0	1	36	2	14	51
2009	96	45	16	2060	3	2	137	18	48	119
2010	174	106	51	4513	3	6	242	32	98	281
2011	256	365	49	7379	9	11	613	82	188	467
2012	598	963	99	14003	28	25	914	106	335	960
2013	1223	1915	160	24905	72	35	1470	191	475	1603
2014	1601	4212	220	37221	166	79	2017	331	787	2683
2015	2875	6764	389	59196	192	90	3064	526	1311	5251
2016	4268	8316	566	82639	184	110	3743	730	1939	5467
2017	6243	9617	890	128006	258	111	5829	1099	2216	7271
2018	7357	10240	1134	131833	217	111	6016	1094	3896	9118
2019	8547	9655	1584	145459	245	113	7734	1166	4416	10232

Table A.8: Values for Figure 3.9: POS by copyright symbol

Year	AdjP	AdvP	INTJ	NP	NP+Adv	NP+P	Other	PP	S	VP
2006	0	0	0	10	0	0	0	0	0	0
2007	0	1	0	58	0	0	3	1	0	7
2008	6	3	1	141	0	0	12	1	0	6
2009	17	6	3	260	0	0	27	9	9	36
2010	22	12	2	541	1	0	60	21	8	43
2011	65	15	8	1007	1	2	84	18	11	107
2012	64	23	6	1473	2	1	126	37	27	153
2013	121	47	17	2335	3	3	197	61	40	214
2014	124	69	24	2862	5	5	238	71	52	261
2015	242	133	32	15288	5	5	345	99	93	467
2016	347	251	45	7383	6	7	455	177	114	1232
2017	379	180	84	8919	15	10	651	238	166	1278
2018	307	187	56	7341	9	6	552	221	113	1324
2019	387	239	91	9169	7	15	758	363	192	1286

Table A.9: Values for Figure 3.10: POS by registered trademark

Year	AdjP	AdvP	INTJ	NP	NP+Adv	NP+P	Other	PP	S	VP
2006	0	0	0	41	0	0	5	0	0	1
2007	4	0	1	240	0	0	12	3	2	7
2008	16	7	2	578	0	0	22	7	4	24
2009	42	18	11	2492	4	3	80	26	30	74
2010	88	19	9	2951	2	10	125	29	56	138
2011	118	45	18	5703	7	7	187	32	62	197
2012	159	71	32	10315	8	20	374	53	114	276
2013	300	105	62	16817	18	22	675	115	206	545
2014	379	209	82	22136	14	36	950	139	404	770
2015	537	236	106	29146	26	67	1019	188	1176	1198
2016	967	373	138	43855	30	146	1375	286	2251	1326
2017	873	366	185	40021	23	159	1498	334	730	1718
2018	874	405	223	43346	35	117	1746	518	1148	1829
2019	1014	519	268	53417	22	79	1765	1565	861	2448

Table A.10: Values for Figure 3.19: Case study part of speech distribution by subreddit

subreddit	AdjP	AdvP	INTJ	NP	NP+Adv	NP+P	Other	PP	S	VP
2b2t	125	80	23	3865	0	3	118	36	40	175
COMPLETEANARCHY	100	69	35	1282	5	2	175	18	96	256
transgendercirclejerk	52	45	14	797	3	0	129	10	19	90

Table A.11: Values for Figure 6.6: Survey 1 function by social media site

Social Media Site	Emphasis	Familiarity	Humor	Other	Quintessential	Quotation	Sarcasm
Facebook	5	8	0	14	0	0	3
Instagram	35	48	8	31	11	5	34
Pinterest	19	20	8	12	2	4	28
Reddit	0	11	1	2	2	0	8
Snapchat	29	30	7	20	9	5	27
Tik Tok	8	27	3	18	9	0	17
Tumblr	0	11	1	8	2	0	8
Twitter	24	38	6	23	7	1	26
YouTube	30	46	12	23	12	5	62

Table A.12: Values for Figure 6.11: Survey 2 rating frequency by function

Function	1 (Lowest)	2	3	4	5 (Highest)	Weighted Mean
attitude	156	58	71	87	108	274.6
stereotype	182	68	81	78	71	245.6
upscaling	232	53	46	56	93	233
familiarity	225	62	73	62	58	221.2
sarcasm	264	51	50	53	62	207.6
quotation	315	58	55	30	22	165.2

Table A.13: Values for Figure 6.12: Survey 2 rating frequency by function (TM)

Function	1 (Lowest)	2	3	4	5 (Highest)	Weighted Mean
attitude	38	13	22	23	24	68.4
stereotype	45	16	24	21	14	60.6
familiarity	56	14	15	17	18	57.4
upscaling	60	16	9	16	19	55.6
sarcasm	64	11	13	19	13	53.2
quotation	69	22	14	7	8	44.6

Table A.14: Values for Figure 6.13: Survey 2 rating frequency by function (HT)

Function	**1 (Lowest)**	**2**	**3**	**4**	**5 (Highest)**	**Weighted Mean**
attitude	37	15	10	27	31	72
familiarity	47	15	18	21	19	62
upscaling	57	10	15	11	27	60.2
stereotype	48	22	18	12	20	58.8
sarcasm	72	10	14	9	15	49
quotation	79	16	15	3	7	40.6

Table A.15: Values for Figure 6.14: Survey 2 rating frequency by function (R)

Function	**1 (Lowest)**	**2**	**3**	**4**	**5 (Highest)**	**Weighted Mean**
attitude	44	18	14	19	25	64.6
stereotype	45	14	23	21	17	62.2
upscaling	57	12	10	19	22	59.4
sarcasm	68	16	8	11	17	50.6
familiarity	62	19	20	11	8	48.8
quotation	86	9	13	9	3	38.8

Table A.16: Values for Figure 6.15 Survey 2 rating frequency by function (C)

Function	**1 (Lowest)**	**2**	**3**	**4**	**5 (Highest)**	**Weighted Mean**
attitude	37	12	25	18	28	69.6
stereotype	44	16	16	24	20	64
upscaling	58	15	12	10	25	57.8
sarcasm	60	14	15	14	17	54.8
familiarity	60	14	20	13	13	53
quotation	81	11	13	11	4	41.2

Table A.17: Values for Figure 7.2: Distribution of author AO frequency for regression task

Target	**Hashtag**	**Copyright**	**Registered**	**Trademark**
0	66920	239032	218397	53421
1	28342	24681	36932	146306
2-5	55012	5147	12638	58377
6-10	32028	384	1030	7352
11-20	31186	102	274	2726
21-50	31070	47	115	999
51-100	13456	14	20	160
101-500	10240	9	11	69
501+	1165	3	2	9

Table A.18: Values for Figure 7.4: Author min and max year on Reddit for regression task

Year	Min	Max
2005	20	0
2006	471	5
2007	1216	24
2008	2037	49
2009	3822	131
2010	7642	245
2011	16725	447
2012	24427	738
2013	25600	1204
2014	27715	2072
2015	27985	4657
2016	28244	6454
2017	25720	9449
2018	30753	12929
2019	25982	17778
2020	16956	32327
2021	4104	180910

Table A.19: Values for Figure 7.5: Distribution of subreddit AO frequency for regression task

Target	Hashtag	Copyright	Registered	Trademark
0	1441	14517	13142	3204
1	465	3496	3441	7256
2-5	1051	2227	2838	5483
6-10	805	529	803	1773
11-20	1212	358	567	1335
21-50	2184	248	426	1152
51-100	2090	70	157	547
101-500	5585	50	112	600
501+	6669	7	16	152

Table A.20: Values for Figure 7.7: Subreddit min and max years on Reddit for regression task

Year	**Min**	**Max**
2005	1	0
2006	8	0
2007	11	0
2008	537	1
2009	809	1
2010	1393	6
2011	2110	7
2012	2368	15
2013	1903	37
2014	1844	57
2015	1870	127
2016	1725	169
2017	1554	253
2018	1867	435
2019	1791	744
2020	1333	1100
2021	378	18550

Table A.21: Values for Figure 7.11: Author min and max years for classification task

Year	**Min**	**Max**
2005	138	0
2006	3555	191
2007	9848	1006
2008	19669	2339
2009	49755	6298
2010	124857	16583
2011	351013	46097
2012	648908	113117
2013	753300	192151
2014	851206	279243
2015	945797	417619
2016	1123308	488466
2017	1201320	610748
2018	2005982	781906
2019	2715798	1199405
2020	3116002	2582093
2021	905415	8088609

Table A.22: Values for Figure 7.13: AO frequency for top subreddit features from random forest classifier

Subreddit	**AO Relative Frequency**	**Hashtag**	**Copyright**	**Trademark**	**Registered**
ShitRedditSays	0.07%	19574	31	618	65
Gamingcirclejerk	0.02%	19525	123	5033	226
HailCorporate	0.02%	861	8	246	34
RandomActsOfGaming	0.04%	6665	0	366	177
copypasta	0.04%	34194	402	3107	1033
niceguys	0.02%	12249	68	4256	90
BreadTube	0.03%	7639	20	261	33
MEAN	0.05%	348.554	3.301	46.582	5.062

Table A.23: Values for Figure 7.14: AO ranking for top subreddit features from random forest classifier

Subreddit	**Hashtag Rank**	**Copyright Rank**	**Trademark Rank**	**Registered Rank**	**Mean Rank**
copypasta	2	2	5	1	2.5
Gamingcirclejerk	4	4	2	4	3.5
RandomActsOfGaming	10	652	19	5	171.5
niceguys	6	7	3	12	7
ShitRedditSays	3	16.5	11	17	11.875
HailCorporate	49	47	27	24	36.75
BreadTube	9	23	25	25	20.5

APPENDIX B: PART OF SPEECH CLUSTER ANNOTATION SCHEME

This appendix contains the detailed annotation scheme used to assign POS tag combinations into POS clusters described in chapter 3.

Table B.1: Annotation scheme for assigning POS clusters

POS Cluster	Frequency	Example	Code
ADJ_NOUN	215091	Awesome Dad	NP
PROPN_NOUN	132915	Hamilton problems	NP
PRON_ADV	128442	Me too	S
ADJ_PROPN	126060	Frisky Friday	NP
NOUN_ADP	123005	Thumbs up	NP+P
DET_NOUN	97615	The network	NP
VERB_DET_NOUN	81010	Save the trees	VP
VERB_NOUN	74894	Show tooltip	VP
DET_ADJ_NUM	74438	The smart one	NP
NOUN_PROPN	70729	Music Monday	NP
PROPN_PROPN_NOUN	51088	New York nerds	NP
VERB_PROPN	45648	End hunger	S
ADJ_NOUN_NOUN	45329	First world problems	NP
PART_DET_NOUN	36233	Not all men	NP
PROPN_VERB	32695	Netflix helps	S
INTJ_PROPN	31731	Thanks Obama	S
NOUN_VERB	27316	Shots fired	S
DET_NOUN_NOUN	21652	My zip code	NP
VERB_ADV	21341	Follow back	VP
PROPN_ADP	19619	Redbull Out	NP+P
ADV_PROPN	19419	Literally Bi	AdvP
PROPN_NOUN_NOUN	19002	Reddit feature request	NP
ADV_ADJ	18980	Forever alone	AdvP
NOUN_ADP_NOUN	18923	War on gardens	NP
ADJ_NOUN_PUNCT	18486	Black lives matter	S
ADV_VERB	18087	Always scheming	AdvP
PROPN_ADP_PROPN	16954	Donald for Spiderman	NP
VERB_PRON	16753	Believe me	VP
DET_PROPN	16633	The Stripes	NP

ADP_PROPN	15564	On Wisconsin	PP
VERB_ADJ_NOUN	14727	Stop sneaky bullshit	VP
VERB_ADP_PROPN	14552	Pray for London	VP
DET_ADJ_NOUN	14169	The good life	NP
X_PROPN	13746	Xx fitness	Other
VERB_ADJ	13660	Feeling depressed	VP
VERB_ADP	13462	Start over	VP
VERB_DET_PROPN	12365	Ask the Sox	VP
PROPN_ADJ	12186	Vermont strong	AdjP
PRON_VERB	12037	I believe	S
VERB_PRON_PROPN_VERB_NOUN	11737	Thank you Pewdie Pie	VP
ADJ_ADJ_NOUN	11196	Famous last words	NP
VERB_ADP_NOUN	11188	Run with scissors	VP
NOUN_ADP_PROPN	10938	Pasta with Darryl	NP
VERB_PROPN_PROPN	10258	Making Forrest Gump	VP
PROPN_DET_NOUN	10045	Assad the liar	NP
PROPN_PROPN_VERB	9862	Bear Grylls drinking	S
PROPN_ADV	9653	Joffrey forever	NP+Adv
PRON_VERB_ADP_PRON	9194	We stand with her	S
VERB_NOUN_NOUN	8987	Stop gamer gate	VP
ADP_NOUN	8987	For science	PP
ADV_ADJ_NOUN	7341	Totally worthless rant	NP
ADJ_NOUN_ADV	3445	Eternal torment now	NP+Adv

APPENDIX C: ITEMS FROM SURVEY 1

The following list contains the items from Survey 1. All 15 participants who took the survey saw all 40 items. Items are grouped by operator and whether or not they are genuine or figurative.

Copyright Symbol: Figurative

1. They've got to use a better example than Jurassic Park, which may not have been The Best Thing Ever©, but was absolutely not a bad novel. Quite the contrary, even beyond the eye-opening premise of the possibilities of genetic replication and the wide appeal of "Yay dinosaurs rar!", the novel proved to be a fascinating look at chaos theory and its applications in real life - something that was actually largely absent from the film. Beyond that, it made for a tense thriller, particularly in the later parts of the book focusing on the raptor siege.

2. He wasn't a "True Christian©" according to the Catholic Church and he got thrown to the fire... seems like the point still stands. Of course, your point about it being rare is valid, but all I need to do is bring up the Inquisition... and continue through as many examples as I can find of people being tortured/murdered for worshiping the wrong god (or even the same god in the wrong way).

3. Unfortunately, plans changed© and it turns out I'll be in the bus for most of the show... But thank you for your answer ! Enjoy Mania !

4. Well, the war certainly settled the issue didn't it?

   Violence©- Solving Your Problems Since Cain smacked Able

5. Those people are probably ignorant, hell they might be racist, but they aren't liars nor did they run on a platform of "change" ©

Copyright Symbol: Genuine

6. AA is not allied with any sect, denomination, politics, organization or institution; does not wish to engage in any controversy, neither endorses nor opposes any causes. Our primary purpose is to stay sober and help other alcoholics to achieve sobriety.
   - Copyright © by The A.A. Grapevine, Inc. Quoted by fair use exemption.

7. Good luck!

   http://bible.cc/1_timothy/5-8.htm

   English Standard Version (©2001)

   But if anyone does not provide for his relatives, and especially for members of his household, he has denied the faith and is worse than an unbeliever.

8. Original article: [© 2011 American Society for Nutrition Red meat consumption and risk of type 2 diabetes: 3 cohorts of US adults and an updated meta-analysis](http://www.ajcn.org/content/early/2011/08/10/ajcn.111.018978.abstract?sid=2ebcfebe-79b1-4c1f-8f6b-4c7c13d781b3)

   Conclusion: Our results suggest that red meat consumption, particularly processed red meat, is associated with an increased risk of T2D.

9. The Baha'i World Centre is taking responsibility for the replacement of the eagle, should the original not be found. We are deeply grateful to those friends who have offered to contribute to the replacement of the eagle or even to sculpt a new one. However, in caring for the Guardian's Resting Place the National Assembly is acting as an agent for the Universal House of Justice, which has the ultimate responsibility for decisions relating to this Holy Place. We anxiously await the guidance of the House of Justice. With loving Baha'i greetings. National Spiritual Assembly of the Baha'is of the UK. Barney Leith, Secretary . ©Copyright 2000, National Spiritual Assembly of the Bahá'ís of the United Kingdom

10. This is the Tool Box, a robust list chock-full of helper applications and homepage development tools for both the casual surfer and the hardcore homepage creator. From graphics viewers to cgi-bin files, this is the spot! Mapedit 1.1.2. A WYSIWYG (What you see is what you get) editor for imagemaps, this application lets you easily create imagemaps by drawing rectangles, circles, polygons, etc. directly on top of your image. It then generates an imagemap file. This is shareware. Versions available for X11 and Windows. Copyright © 1994 by Thomas Boutell.

Registered Trademark: Figurative

11. That's just god trying to show you the Truth®

12. I agree with you. He gets a lot of respect from this sub, but the Greater NASCAR Empire® does not seem to.

13. Finally somebody said it! That's what the patriarchy® is trying to say all along!

14. On the one hand I'm glad we get to hear new songs in movie trailers, on the other, *quirky sentimental indie* ® is apparently the new 'feel-good family comedy with heart' music. It works here, but in the Winnie the Pooh trailer, I honestly thought it was a fan mash-up.

15. Hopefully not. In my wildest dream I would like to see Nokia buy it back and ship the Best Phone Ever®, but I doubt Nokia now has the foresight they lacked during the Microsoft takeover. OTOH, I wouldn't exclude some interest in the company by someone in India or China.

Registered Trademark: Genuine

16. Microsoft® Windows® XP with Service Pack 2 (Service Pack 3 recommended) or Windows Vista® Home Premium, Business, Ultimate, or Enterprise with Service Pack 1 (certified for 32-bit Windows XP and Windows Vista)

17. This image was enhanced using Adobe® Photoshop® software.

18. Leap Frog Connected Learning Game System Introducing Leapster2 - the only preschool learning game system that offers personalized insights into what your child is learning. Following on the heels of the best-selling Leapster® learning game system, the new Leapster2 handheld is the next generation of learn-everywhere gaming from LeapFrog. Like the Leapster learning game system, the Leapster2 handheld offers a robust learning experience through built-in tutorials and learning levels that adapt automatically to your child’s pace. Its touch screen and stylus help develop motor skills used in writing, while its compact design makes it easy for kids to play on the go. It’s also compatible with all 30+ Leapster learning games, so kids can practice a wide variety of skills for school as they play and learn with their favorite characters......

19. Campbell's. So many, many reasons it's so...M'm! M'm! Good!®

20. I was eight years old and playing with LEGO®.

Trademark: Figurative

21. A picture of My Favorite President™!? How did you know?

22. Reagan was the president brought to us by GE. As I recall, there really wasn't much for him to officially do except show up at events, hit his marks, and run through his 5"x7" cards on time with a minimum of obvious gaffes. The magic of Capitalism™ handled all the really heavy lifting. There was the time he thought trees caused more pollution than cars, and other times when his eyes would wonder off the script and he would nod about confused until a trusted aide stepped in to steer him back on course.

23. PopCuts: Where it pays to be a condescending hipster™

24. They can still *Just say No!™*

25. Seriously. As soon who grew up during the Vietnam Police Action and protested the draft, I now see I was completely wrong. We need to return to a military draft. If we're going to have wars *everyone's* kid needs to be fair game for cannon fodder, not just those of the poor. I would go so far as to eliminate military recruitment and volunteering. The military may be a necessary evil, but we shouldn't necessarily admit those who are too gung-ho about it.

    They were obviously in a hurry to get everything set up by the end of a single movie, and it shows. Still, it's not a Big Deal™ if y'all look at how well both sides can argue. It's fiction and amounts to a detail in an otherwise very entertaining film.

Trademark: Genuine

26. That's not quite correct: they've approved the Reb-A extract of Stevia, not the plant itself, but presumably Truvia™ and PureVia™ based sweetener packs should be available soon alongside Splenda and the rest.

27. PB & Nutella™.

    Love Nutella! Goes great with PB, for sure.

28. Operating System: Windows Vista™ Home Premium (6.0, Build 6001) Service Pack 1
    - System Manufacturer: ASUSTeK Computer Inc.

29. It's not really all about how still the camera is, it's all about how well you can Photoshop™ it.

30. That article really made me want to eat some McDonald's™ fries.

Hashtag: Figurative

31. It's funny the IWC complains "We want The Attitude Era back!" yet when something that would have happened in The Attitude Era goes down they piss and moan or pile on the faux outrage for #wokepoints. This is professional wrestling. It is not, was not, and never will be politically correct. The Attitude Era sure as hell wasn't.

32. As much of Africa is on the rise, the situations in Burundi and Central African Republic remain very #Bad

33. No it's #TheBest behavior!

34. Can we just agree that Trump and McCain both suck and that what Trump is doing is also still petty and childish? Every person antithetical to Trump is suddenly branded as #Resistance heroes and it's tiring.

35. Ah the socialist democrats of reddit. Bail is terrible. Bail bonds are terrible.

    Ok your sister had the shit beat out of her. Now since bail is wrong. The offender isnt in jail. Now your sister is in hiding because an order of protection will not keep her safe

    Or lets get rid of bail bonds. Thats a great idea. Those poor folks who actually get thrown into jail wont be able to bond out. So now they must put up the full bail amount. Which obviously they cant do.

    Both of these things will happen if those idiots who dont think things through ... like ya'll on reddit ... change the system.

    But im sure #feelings are more important.

Hashtag: Genuine

36. I like to do #antiinstagram posts. Like when I get lost and drive down W.VA roads that are 90% potholes or cry because of the CRUSHING SILENCE AND GOD DAMN LEAKS... really breaks up the romantic view

37. I agree, and there are people who, I suppose are from other churches or other parts of the world, that are still posting using #wakeupolive to say that they’re not giving up, she will rise, etc...

38. Think we can get #RaiseTheArmada trending?

39. Antonio Paladino postet a photo of a tattooed guy one month ago with the hashtags #capetown and #tatoo. I don't think this is an coincidence.

40. Haha, me too! I was searching by the #lisse hashtag. I want to see her other work. :)

APPENDIX D: ITEMS FROM SURVEY 2

The following list contains the experimental items from Survey 2. There were four versions of the survey, varying by which AO was used in which item. Each of the versions is presented below.

Version A

1. He was High on Friendship™!

2. That, my friend, is what we call #TheAmericanWay

3. Get yourself some better speakers or maybe a nice pair of headphones. Seriously. At least try it for Science®. Details pop out of the noise; order from chaos. Or, just keep listening to nickelback, I don't care.

4. Finally some sense. I guess Change© is coming.

5. Whoa. I had a bit of disorientation at first, kind of like a Escher Moment™.

6. But actually looking at real studies and listening to actual trans people would cause these people to have empathy and discover a new part of the world around them, and acknowledging and respecting trans people will topple them from the #Hierarchy and we can't have that

7. This guy knows The Rules®.

8. but you get The Best Healthcare in The World© in America, so it's worth anything, right?

9. I give out Cool Points™ all the time. But, they can be taken away just as quickly. It's fun to give the look of disapproval and say, "... you're minus three. I expect a little effort on your part."

10. Now I have to go watch it. #again

11. You were one of the first Europeans to be experiencing the madness of MLS After Dark® upon waking up.

12. Ok, funny story (to me anyway!) My kids used to watched this around ages 11 and 7. I actually really love shows like this, but the first time I saw naruto it was this scene and it just struck me as so funny. My kids were irritated because I was crying I was laughing so hard, but they got REALLY© pissed when I said "They battle by playing patty-cakes?

Version B

1. He was High on Friendship©!

2. That, my friend, is what we call The American Way™

3. Get yourself some better speakers or maybe a nice pair of headphones. Seriously. At least try it for #Science. Details pop out of the noise; order from chaos. Or, just keep listening to nickelback, I don't care.

4. Finally some sense. I guess Change® is coming.

5. Whoa. I had a bit of disorientation at first, kind of like a Escher Moment©.

6. But actually looking at real studies and listening to actual trans people would cause these people to have empathy and discover a new part of the world around them, and acknowledging and respecting trans people will topple them from the Hierarchy™ and we can't have that

7. This guy knows #TheRules.

8. but you get The Best Healthcare in The World® in America, so it's worth anything, right?

9. I give out Cool Points© all the time. But, they can be taken away just as quickly. It's fun to give the look of disapproval and say, "... you're minus three. I expect a little effort on your part."

10. Now I have to go watch it. again™

11. You were one of the first Europeans to be experiencing the madness of #MLSAfterDark upon waking up.

12. Ok, funny story (to me anyway!) My kids used to watched this around ages 11 and 7. I actually really love shows like this, but the first time I saw naruto it was this scene and it just struck me as so funny. My kids were irritated because I was crying I was laughing so hard, but they got REALLY® pissed when I said "They battle by playing patty-cakes?!

Version C

1. He was High on Friendship®!

2. That, my friend, is what we call The American Way©

3. Get yourself some better speakers or maybe a nice pair of headphones. Seriously. At least try it for Science™. Details pop out of the noise; order from chaos. Or, just keep listening to nickelback, I don't care.

4. Finally some sense. I guess #Change is coming.

5. Whoa. I had a bit of disorientation at first, kind of like a Escher Moment®.

6. But actually looking at real studies and listening to actual trans people would cause these people to have empathy and discover a new part of the world around them, and acknowledging and respecting trans people will topple them from the Hierarchy© and we can't have that

7. This guy knows The Rules™.

8. but you get #TheBestHealthcareinTheWorld in America, so it's worth anything, right?

9. I give out Cool Points® all the time. But, they can be taken away just as quickly. It's fun to give the look of disapproval and say, "... you're minus three. I expect a little effort on your part."

10. Now I have to go watch it. again©

11. You were one of the first Europeans to be experiencing the madness of MLS After Dark™ upon waking up.

12. Ok, funny story (to me anyway!) My kids used to watched this around ages 11 and 7. I actually really love shows like this, but the first time I saw naruto it was this scene and it just struck me as so funny. My kids were irritated because I was crying I was laughing so hard, but they got #REALLY pissed when I said "They battle by playing patty-cakes?!

Version D

1. He was #HighonFriendship!

2. That, my friend, is what we call The American Way®

3. Get yourself some better speakers or maybe a nice pair of headphones. Seriously. At least try it for Science©. Details pop out of the noise; order from chaos. Or, just keep listening to nickelback, I don't care.

4. Finally some sense. I guess Change™ is coming.

5. Whoa. I had a bit of disorientation at first, kind of like a #EscherMoment.

6. But actually looking at real studies and listening to actual trans people would cause these people to have empathy and discover a new part of the world around them, and acknowledging and respecting trans people will topple them from the Hierarchy® and we can't have that

7. This guy knows The Rules©.

8. but you get The Best Healthcare in The World™ in America, so it's worth anything, right?

9. I give out #CoolPoints all the time. But, they can be taken away just as quickly. It's fun to give the look of disapproval and say, "... you're minus three. I expect a little effort on your part."

10. Now I have to go watch it. again®

11. You were one of the first Europeans to be experiencing the madness of MLS After Dark© upon waking up.

12. Ok, funny story (to me anyway!) My kids used to watched this around ages 11 and 7. I actually really love shows like this, but the first time I saw naruto it was this scene and it just struck me as so funny. My kids were irritated because I was crying I was laughing so hard, but they got REALLY™ pissed when I said "They battle by playing patty-cakes?!

APPENDIX E: DETAILED STATISTICAL RESULTS FROM SURVEY 2

This appendix contains the detailed statistical results from Survey 2 which were summarized in chapter 6.

Table E.1: Ordinal logistic regression: trademark model

	Estimate	Std. Error	p
Familiarity	-0.7933	0.4084	0.0521
Quotation	-1.8009	0.5813	0.00195
Sarcasm	-1.168	0.4973	0.01885
Stereotype	-0.4565	0.3447	0.18538
Upscaling	-0.9517	0.382	0.01272

Table E.2: Emmeans pairwise comparisons: trademark model

contrast		estimate	SE	df	z.ratio	p
attitude	familiarity	0.793	0.408	Inf	1.942	0.3762
attitude	quotation	1.801	0.581	Inf	3.098	0.0239
attitude	sarcasm	1.168	0.497	Inf	2.349	0.1748
attitude	stereotype	0.456	0.345	Inf	1.324	0.7716
attitude	upscaling	0.952	0.382	Inf	2.491	0.1265
familiarity	quotation	1.008	0.479	Inf	2.101	0.2864
familiarity	sarcasm	0.375	0.573	Inf	0.654	0.9867
familiarity	stereotype	-0.337	0.359	Inf	-0.939	0.9365
familiarity	upscaling	0.158	0.41	Inf	0.387	0.9989
quotation	sarcasm	-0.633	0.682	Inf	-0.929	0.9393
quotation	stereotype	-1.344	0.575	Inf	-2.337	0.1792
quotation	upscaling	-0.849	0.498	Inf	-1.706	0.5278
sarcasm	stereotype	-0.712	0.454	Inf	-1.566	0.6213
sarcasm	upscaling	-0.216	0.598	Inf	-0.362	0.9992
stereotype	upscaling	0.495	0.404	Inf	1.227	0.8238

Table E.3: Ordinal logistic regression: hashtag model

	Estimate	Std. Error	p
Familiarity	-0.6804	0.5526	0.21817
Quotation	-2.5249	0.6081	3.30E-05
Sarcasm	-1.8168	0.6557	0.00559
Stereotype	-0.8284	0.4351	0.05691
Upscaling	-0.8888	0.5003	0.07561

Table E.4: Emmeans pairwise comparisons: hashtag model

contrast		estimate	SE	df	z.ratio	p
attitude	familiarity	0.6804	0.553	Inf	1.231	0.8215
attitude	quotation	2.5249	0.608	Inf	4.152	0.0005
attitude	sarcasm	1.8168	0.656	Inf	2.771	0.0622
attitude	stereotype	0.8284	0.435	Inf	1.904	0.3996
attitude	upscaling	0.8888	0.5	Inf	1.777	0.4809
familiarity	quotation	1.8444	0.521	Inf	3.541	0.0054
familiarity	sarcasm	1.1364	0.707	Inf	1.607	0.5938
familiarity	stereotype	0.148	0.402	Inf	0.368	0.9991
familiarity	upscaling	0.2084	0.484	Inf	0.431	0.9981
quotation	sarcasm	-0.708	0.603	Inf	-1.174	0.8493
quotation	stereotype	-1.6964	0.531	Inf	-3.193	0.0177
quotation	upscaling	-1.636	0.682	Inf	-2.4	0.1561
sarcasm	stereotype	-0.9884	0.635	Inf	-1.557	0.6272
sarcasm	upscaling	-0.928	0.737	Inf	-1.26	0.807
stereotype	upscaling	0.0604	0.471	Inf	0.128	1

Table E.5: Ordinal logistic regression: registered trademark model

	Estimate	Std. Error	p
Familiarity	-0.9685	0.3989	0.015178
Quotation	-2.1756	0.5699	0.000135
Sarcasm	-1.0842	0.4714	0.021441
Stereotype	-0.1713	0.325	0.598222
Upscaling	-0.4529	0.4762	0.341653

Table E.6: Emmeans pairwise comparisons: registered trademark model

contrast		estimate	SE	df	z.ratio	p
attitude	familiarity	0.968	0.399	Inf	2.428	0.1465
attitude	quotation	2.176	0.57	Inf	3.817	0.0019
attitude	sarcasm	1.084	0.471	Inf	2.3	0.1938
attitude	stereotype	0.171	0.325	Inf	0.527	0.9951
attitude	upscaling	0.453	0.476	Inf	0.951	0.9331
familiarity	quotation	1.207	0.454	Inf	2.658	0.0839
familiarity	sarcasm	0.116	0.589	Inf	0.197	1
familiarity	stereotype	-0.797	0.306	Inf	-2.606	0.0957
familiarity	upscaling	-0.516	0.484	Inf	-1.066	0.895
quotation	sarcasm	-1.091	0.667	Inf	-1.637	0.5737
quotation	stereotype	-2.004	0.502	Inf	-3.993	0.0009
quotation	upscaling	-1.723	0.735	Inf	-2.343	0.1768
sarcasm	stereotype	-0.913	0.523	Inf	-1.745	0.5021
sarcasm	upscaling	-0.631	0.71	Inf	-0.889	0.9493
stereotype	upscaling	0.282	0.449	Inf	0.627	0.9891

Table E.7: Ordinal logistic regression: copyright model

	Estimate	Std. Error	p
Familiarity	-1.0357	0.4412	0.0189
Quotation	-2.101	0.496	2.27E-05
Sarcasm	-0.9213	0.4604	0.0454
Stereotype	-0.2871	0.4264	0.5007
Upscaling	-0.7401	0.3955	0.0613

Table E.8: Emmeans pairwise comparisons: copyright model

contrast		estimate	SE	df	z.ratio	p
attitude	familiarity	1.036	0.441	Inf	2.347	0.1753
attitude	quotation	2.101	0.496	Inf	4.236	0.0003
attitude	sarcasm	0.921	0.46	Inf	2.001	0.3416
attitude	stereotype	0.287	0.426	Inf	0.673	0.9849
attitude	upscaling	0.74	0.396	Inf	1.871	0.4199
familiarity	quotation	1.065	0.387	Inf	2.755	0.0649
familiarity	sarcasm	-0.114	0.536	Inf	-0.214	0.9999
familiarity	stereotype	-0.749	0.293	Inf	-2.558	0.1079
familiarity	upscaling	-0.296	0.47	Inf	-0.629	0.9889
quotation	sarcasm	-1.18	0.617	Inf	-1.914	0.3937
quotation	stereotype	-1.814	0.426	Inf	-4.26	0.0003
quotation	upscaling	-1.361	0.491	Inf	-2.773	0.0618
sarcasm	stereotype	-0.634	0.467	Inf	-1.357	0.7527
sarcasm	upscaling	-0.181	0.586	Inf	-0.309	0.9996
stereotype	upscaling	0.453	0.473	Inf	0.957	0.9314

Table E.9: Ordinal logistic regression: attitude model

	Estimate	Std. Error	p
symbol_cleanC	-0.1607	0.3204	0.6159
symbol_cleanR	-0.5116	0.2765	0.0643
symbol_cleanTM	-0.2046	0.2929	0.4848

Table E.10: Emmeans pairwise comparisons: attitude model

contrast		estimate	SE	df	z.ratio	p
#	C	0.1607	0.32	Inf	0.502	0.9587
#	R	0.5116	0.277	Inf	1.85	0.2498
#	TM	0.2046	0.293	Inf	0.699	0.8976
C	R	0.3509	0.296	Inf	1.186	0.6358
C	TM	0.0439	0.297	Inf	0.148	0.9989
R	TM	-0.307	0.285	Inf	-1.077	0.7038

Table E.11: Ordinal logistic regression: familiarity model

	Estimate	Std. Error	p
C	-0.6871	0.3246	0.0343
R	-0.8633	0.349	0.0134
TM	-0.5243	0.3882	0.1768

Table E.12: Emmeans pairwise comparisons: familiarity model

contrast		estimate	SE	df	z.ratio	p
#	C	0.687	0.325	Inf	2.117	0.1477
#	R	0.863	0.349	Inf	2.473	0.0641
#	TM	0.524	0.388	Inf	1.351	0.5306
C	R	0.176	0.327	Inf	0.539	0.9495
C	TM	-0.163	0.366	Inf	-0.445	0.9706
R	TM	-0.339	0.381	Inf	-0.891	0.8096

Table E.13: Ordinal logistic regression: quotation model

	Estimate	Std. Error	p
C	0.27532	0.61661	0.655
R	0.03355	0.6325	0.958
TM	0.74981	0.55478	0.177

Table E.14: Emmeans pairwise comparisons: quotation model

contrast		estimate	SE	df	z.ratio	p
#	C	-0.2753	0.617	Inf	-0.447	0.9703
#	R	-0.0336	0.633	Inf	-0.053	0.9999
#	TM	-0.7498	0.555	Inf	-1.352	0.53
C	R	0.2418	0.534	Inf	0.453	0.969
C	TM	-0.4745	0.418	Inf	-1.134	0.6686
R	TM	-0.7163	0.509	Inf	-1.407	0.4949

Table E.15: Ordinal logistic regression: sarcasm model

	Estimate	Std. Error	p
C	0.5757	0.3506	0.101
R	0.2418	0.3721	0.516
TM	0.3616	0.4059	0.373

Table E.16: Emmeans pairwise comparisons: sarcasm model

contrast		estimate	SE	df	z.ratio	p
#	C	-0.576	0.351	Inf	-1.642	0.3551
#	R	-0.242	0.372	Inf	-0.65	0.9157
#	TM	-0.362	0.406	Inf	-0.891	0.8096
C	R	0.334	0.339	Inf	0.984	0.7585
C	TM	0.214	0.319	Inf	0.672	0.9077
R	TM	-0.12	0.344	Inf	-0.348	0.9855

Table E.17: Ordinal logistic regression: stereotype model

	Estimate	Std. Error	p
C	0.26164	0.28512	0.359
R	0.18666	0.34289	0.586
TM	0.05231	0.28899	0.856

Table E.18: Emmeans pairwise comparisons: stereotype model

contrast		estimate	SE	df	z.ratio	p
#	C	-0.2616	0.285	Inf	-0.918	0.7954
#	R	-0.1867	0.343	Inf	-0.544	0.9481
#	TM	-0.0523	0.289	Inf	-0.181	0.9979
C	R	0.075	0.301	Inf	0.249	0.9946
C	TM	0.2093	0.256	Inf	0.819	0.8456
R	TM	0.1343	0.279	Inf	0.482	0.9631

Table E.19: Ordinal logistic regression: upscaling model

	Estimate	Std. Error	p
C	-0.11367	0.31837	0.721
R	-0.08173	0.341	0.811
TM	-0.27176	0.32995	0.41

Table E.20: Emmeans pairwise comparisons: upscaling model

contrast		estimate	SE	df	z.ratio	p
#	C	0.1137	0.318	Inf	0.357	0.9844
#	R	0.0817	0.341	Inf	0.24	0.9952
#	TM	0.2718	0.33	Inf	0.824	0.8433
C	R	-0.0319	0.293	Inf	-0.109	0.9995
C	TM	0.1581	0.301	Inf	0.526	0.9529
R	TM	0.19	0.32	Inf	0.594	0.934

REFERENCES

Androutsopoulos, J. (2011). From variation to heteroglossia in the study of computer-mediated discourse. Digital discourse: Language in the new media, 277-298.

Angouri, J. (2015). Online communities and communities of practice. In The Routledge handbook of language and digital communication (pp. 323-338). Routledge.

Argamon, S., Koppel, M., Pennebaker, J. W., & Schler, J. (2007). Mining the blogosphere: Age, gender and the varieties of self-expression. First Monday, 12(9).

Auer, P. (1992). Introduction: John Gumperz’ approach to contextualization. In The contextualization of language (p. 1). John Benjamins.

Baker, P., Gabrielatos, C., Khosravinik, M., Krzyżanowski, M., McEnery, T., & Wodak, R. (2008). A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. Discourse & society, 19(3), 273-306.

Baron, N. S. (2004). See you online: Gender issues in college student use of instant messaging. Journal of language and social psychology, 23(4), 397-423.

Baron, N. S. (2009). The myth of impoverished signal: Dispelling the spoken language fallacy for emoticons in online communication. Electronic Emotion: The Mediation of Emotion via Information and Communication Technologies. Bern: Peter Lang, 107-135.

Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., & Blackburn, J. (2020, May). The pushshift reddit dataset. In Proceedings of the international AAAI conference on web and social media (Vol. 14, pp. 830-839).

Benamara, F., Inkpen, D., & Taboada, M. (2018). Introduction to the special issue on language in social media: Exploiting discourse and other contextual information. Computational Linguistics, 44(4), 663-681.

Biber, D., & Finegan, E. (1989). Styles of stance in English: Lexical and grammatical marking of evidentiality and affect. Text-interdisciplinary journal for the study of discourse, 9(1), 93-124.

Bott, L., & Noveck, I. A. (2004). Some utterances are underinformative: The onset and time course of scalar inferences. Journal of memory and language, 51(3), 437-457.

Brennan, S. E. (1998). The grounding problem in conversations with and through computers. Social and cognitive approaches to interpersonal communication, 201-225.

Browning, Darcey. 2017. #TwitterDiscourseMarkers: A corpora based study of the pragmatic functions of hashtags. University of Texas, Arlington, PhD dissertation.

Bruns, A., & Burgess, J. E. (2011, August). The use of Twitter hashtags in the formation of ad hoc publics. In Proceedings of the 6th European Consortium for Political Research (ECPR) General Conference 2011.

Chen, A. C. H. (2020). Words, constructions and corpora: Network representations of constructional semantics for Mandarin space particles. Corpus Linguistics and Linguistic Theory.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078.

Cirillo, L. (2019). The pragmatics of air quotes in English academic presentations. Journal of Pragmatics, 142, 1-15.

Collister, L. B. (2011). *-repair in Online Discourse. Journal of Pragmatics, 43(3), 918-921.

Colston, H. L. (1997). Salting a wound or sugaring a pill: The pragmatic functions of ironic criticism. Discourse processes, 23(1), 25-45.

Crystal, D. (2001). Language and the Internet. Cambridge University Press.

Crystal, D. (2014). Stylistic profiling. In English Corpus Linguistics (pp. 233-250). Routledge.

Daft, R. L., & Lengel, R. H. (1983). Information richness. A new approach to managerial behavior and organization design (No. TR-ONR-DG-02). Texas A and M Univ College Station Coll of Business Administration.

Davies, B. (2005). Communities of practice: Legitimacy not choice. Journal of Sociolinguistics, 9(4), 557-581.

Deumert, A. (2015). Linguistics and social media. In The Routledge Handbook of Linguistics (pp. 577-589). Routledge.

Deumert, A., & Masinyana, S. O. (2008). Mobile language choices—The use of English and isiXhosa in text messages (SMS): Evidence from a bilingual South African sample. English World-Wide, 29(2), 117-147.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Dews, S., Kaplan, J., & Winner, E. (1995). Why not say it directly? The social functions of irony. Discourse processes, 19(3), 347-367.

De Winter, J. C., Gosling, S. D., & Potter, J. (2016). Comparing the Pearson and Spearman correlation coefficients across distributions and sample sizes: A tutorial using simulations and empirical data. Psychological methods, 21(3), 273.

Dresner, E., & Herring, S. C. (2010). Functions of the nonverbal in CMC: Emoticons and illocutionary force. Communication theory, 20(3), 249-268.

Du Bois, J. W. (2007). The stance triangle. Stancetaking in discourse: Subjectivity, evaluation, interaction, 164(3), 139-182.

Evison, J. (2010). What are the basics of analysing a corpus?. In The Routledge handbook of corpus linguistics (pp. 122-135). Routledge.

Fernandes, S., & Bernardino, J. (2015, July). What is bigquery?. In Proceedings of the 19th International Database Engineering & Applications Symposium (pp. 202-203).

Ferrara, K., Brunner, H., & Whittemore, G. (1991). Interactive written discourse as an emergent register. Written communication, 8(1), 8-34. Chicago

Fitzsimmons-Doolan, S. (2015). Applying corpus linguistics to language policy. Research methods in language policy and planning: A practical guide, 107-117.

Fleckenstein, K. (2019). "Well I Don't like Abortion' Well Then Don't Have One": A Corpus-Assisted Discourse Analysis of the Stance Functions of Some Discourse Markers in Mediated Abortion Debate. University of Texas, Arlington, PhD dissertation.

Fox, A. B., Bukatko, D., Hallahan, M., & Crawford, M. (2007). The medium makes a difference: Gender similarities and differences in instant messaging. Journal of Language and Social Psychology, 26(4), 389-397. Chicago

Gibbs Jr, R. W., & Colston, H. L. (2007). The future of irony studies.

Grice, H. P. (1975). Logic and conversation. In Speech acts (pp. 41-58). Brill.

Grodner, D. J., Klein, N. M., Carbary, K. M., & Tanenhaus, M. K. (2010). "Some," and possibly all, scalar inferences are not delayed: Evidence for immediate pragmatic enrichment. Cognition, 116(1), 42-55.

Goswami, S., Sarkar, S., & Rustagi, M. (2009, March). Stylometric analysis of bloggers' age and gender. In Third international AAAI conference on weblogs and social media.

Gumperz, J. J. (1992). 8 8 Contextualization and. Rethinking context: Language as an interactive phenomenon, (11), 229.

Heath, M. (2017). Interpretations of non-standard capitalization on Twitter (pp. 15-29). LSO Working Papers in Linguistics. Chicago Herring, S. C. (1999, January). Interactional coherence in CMC. In Proceedings of the 32nd Annual Hawaii International Conference on Systems Sciences. 1999. HICSS-32. Abstracts and CD-ROM of Full Papers (pp. 13-pp). IEEE.

Herring, S. C. (1999, January). Interactional coherence in CMC. In Proceedings of the 32nd Annual Hawaii International Conference on Systems Sciences. 1999. HICSS-32. Abstracts and CD-ROM of Full Papers (pp. 13-pp). IEEE.

Herring, S. C., Stein, D., & Virtanen, T. (2013). 1. Introduction to the pragmatics of computermediated communication. In S. Herring, D. Stein, & T. Virtanen (Eds.), Pragmatics of Computer-Mediated Communication. https://doi.org/10.1515/9783110214468.3

Honnibal, M., & Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 7(1), 411-420.

Hunston, S., & Thompson, G. (2001). Evaluation in Text: Authorial Stance and the Construction of Discourse.

Huang, J., Thornton, K. M., & Efthimiadis, E. N. (2010, June). Conversational tagging in twitter. In Proceedings of the 21st ACM conference on Hypertext and hypermedia (pp. 173-178).

Huffaker, D. A., & Calvert, S. L. (2005). Gender, identity, and language use in teenage blogs. Journal of computer-mediated communication, 10(2), JCMC10211.

Jenks, Grant. Wordsegment (1.3.1), July 2018. URL https://github.com/grantjenks/ python-wordsegment.

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759.

Kaiser, E., & Rudin, D. (2020). When faultless disagreement is not so faultless: What widely-held opinions can tell us about subjective adjectives. Proceedings of the Linguistic Society of America, 5(1), 698-707.

Kehoe, A., & Gee, M. (2011). Social Tagging: A new perspective on textual “aboutness”. Studies in Variation, Contacts and Change in English, 6.

Kennedy, C. (2007). Vagueness and grammar: The semantics of relative and absolute gradable adjectives. Linguistics and philosophy, 30(1), 1-45.

Khazraie, M., & Talebzadeh, H. (2020). “Wikipedia does NOT tolerate your babbling!”: impoliteness-induced conflict (resolution) in a polylogal collaborative online community of practice. Journal of Pragmatics, 163, 46-65.

Kiesler, S., Siegel, J., & McGuire, T. W. (1984). Social psychological aspects of computer-mediated communication. American psychologist, 39(10), 1123.

Knight, N. K. (2010). Wrinkling complexity: concepts of identity and affiliation in humour. New discourse on language: Functional perspectives on multimodality, identity, and affiliation. London: Continuum, 35-58.

Kopple, W. J. V. (1985). Some exploratory discourse on metadiscourse. College composition and communication, 82-93.

Kotowski, S. (2021). The semantics of English out-prefixation: A corpus-based investigation. English Language and Linguistics, 25(1), 61-89. doi:10.1017/S1360674319000443

Kronmüller, E., Morisseau, T., & Noveck, I. A. (2014). Show me the pragmatic contribution: a developmental investigation of contrastive inference. Journal of child language, 41(5), 985-1014.

Lenth, R. (2019). Emmeans: Estimated marginal means, aka least-squares means. 2018; R package version 1.3. 1. View Article.

Leuckert, S., & Leuckert, M. (2020). Towards a digital sociolinguistics. Corpus approaches to social media, 98, 15.

Mautner, G. (2009). Corpora and critical discourse analysis. Contemporary corpus linguistics, 32-46.

Markman, K. M., & Oshima, S. (2007, October). Pragmatic play? Some possible functions of English emoticons and Japanese kaomoji in computer-mediated discourse. In Association of Internet Researchers annual conference (Vol. 8).

Martin, J. R., & White, P. R. (2003). The language of evaluation (Vol. 2). Basingstoke: Palgrave Macmillan.

Maynor, N. (1994). 7. The Language of Electronic Mail: Written Speech?. Publication of the American Dialect Society, 78(1), 48-54.

Mauchand, M., Vergis, N., & Pell, M. D. (2020). Irony, prosody, and social impressions of affective stance. Discourse Processes, 57(2), 141-157.

McEnery, T., Xiao, R., & Tono, Y. (2006). Corpus-based language studies: An advanced resource book. Taylor & Francis.

McSweeney, M. (2017, January). Lol! I didn't mean that:"lol" as a marker of the illocutionary force of an utterance. In Poster Presented at the Linguistics Society of America Annual Meeting.

McSweeney, M. A. (2018). The pragmatics of text messaging: making meaning in messages. Routledge.

Meyerhoff, M., & Strycharz, A. (2013). Communities of practice. The handbook of language variation and change, 428-447.

Na'aman, N., Provenza, H., & Montoya, O. (2017, July). Varying linguistic purposes of emoji in (twitter) context. In Proceedings of ACL 2017, Student Research Workshop (pp. 136-141). Chicago

Negrón Goldbarg, R. (2009). Spanish-English codeswitching in email communication. Language@ internet, 6(3).

O'Reilly, T. (2006). Web 2.0 Compact Definition: Trying Again. radar. oreilly. com. Online, 10, 12.

Page, R. (2012). The linguistics of self-branding and micro-celebrity in Twitter: The role of hashtags. Discourse & Communication, 6(2), 181–201.

Pexman, P. M., & Olineck, K. M. (2002). Does sarcasm always sting? Investigating the impact of ironic insults and ironic compliments. Discourse Processes, 33(3), 199-217.

Pexman, P. M., & Zvaigzne, M. T. (2004). Does irony go better with friends?. Metaphor and symbol, 19(2), 143-163.

R/transgendercirclejerk. reddit. (n.d.). Retrieved August 8, 2022, from https://www.reddit.com/r/transgendercirclejerk/

Riloff, E., Qadir, A., Surve, P., De Silva, L., Gilbert, N., & Huang, R. (2013, October). Sarcasm as contrast between a positive sentiment and negative situation. In Proceedings of the 2013 conference on empirical methods in natural language processing (pp. 704-714).

Rossi, L., & Magnani, M. (2012, May). Conversation practices and network structure in Twitter. In Sixth International AAAI Conference on Weblogs and Social Media.

Scott, K. (2018). "Hashtags work everywhere": The pragmatic functions of spoken hashtags. Discourse, context & media, 22, 57-64.

Searle, J. R., & Searle, J. R. (1969). Speech acts: An essay in the philosophy of language (Vol. 626). Cambridge university press.

Shuyo, N. (2010). Language detection library for java.

Siebenhaar, B. (2006). Code choice and code-switching in Swiss-German Internet Relay Chat rooms. Journal of Sociolinguistics, 10(4), 481-506.

Soon™. WoWWiki. (n.d.). Retrieved August 8, 2022, from http://www.wowwiki.com/Soon

Sperber, D., & Wilson, D. (1981). Irony and the use-mention distinction. Philosophy, 3, 143-184.

Stvan, L. S. (2006). Diachronic change in the uses of the discourse markers" why" and" say" in american english. In Corpus linguistics: Applications for the study of English (pp. 61-76). Anubar.

Subtirelu, N. C., & Baker, P. (2017). Corpus-based approaches. In The Routledge handbook of critical discourse studies (pp. 106-119). Routledge.

Tagg, C. (2009). A corpus linguistics study of SMS text messaging (Doctoral dissertation, University of Birmingham).

Tagliamonte, S. A., & Denis, D. (2008). Linguistic ruin? LOL! Instant messaging and teen language. American speech, 83(1), 3-34.

Taylor, C. (2014). Investigating the representation of migrants in the UK and Italian press: A cross-linguistic corpus-assisted discourse analysis. International journal of corpus linguistics, 19(3), 368-400.

Thurlow, C., & Brown, A. (2003). Generation Txt? The sociolinguistics of young people's text-messaging. Discourse analysis online, 1(1), 30.

Uygur-Distexhe, D. (2014). "Lol, mdr and ptdr. An inclusive and gradual approach to discourse markers". En: Cougnon, LA y C. Fairon (eds.), SMS Communication, 239-264.

Vandergriff, I. (2013). Emotive communication online: A contextual analysis of computer-mediated communication (CMC) cues. Journal of Pragmatics, 51, 1-12. Chicago

Werry, C. C. (1996). Internet relay chat. Computer-mediated communication: Linguistic, social and cross-cultural perspectives, 47-63.

Widdowson, H. G. (2000). On the limitations of linguistics applied. Applied linguistics, 21(1), 3-25.

Wikström, P. (2014). # srynotfunny: Communicative functions of hashtags on Twitter. SKY Journal of Linguistics, 27, 127-152.

Williams, E. A. (2021). Pragmatic extension in computer-mediated communication: The case of '#'and '™'. Journal of Pragmatics, 181, 165-179.

Witmer, D. F., & Katzman, S. L. (1997). On-line smiles: Does gender make a difference in the use of graphic accents?. Journal of Computer-mediated communication, 2(4), JCMC244.

Yang, L., Sun, T., Zhang, M., & Mei, Q. (2012, April). We know what@ you# tag: does the dual role affect hashtag adoption?. In Proceedings of the 21st international conference on World Wide Web (pp. 261-270). ACM.

Yates, S. J. (1996). Oral and written linguistic aspects of computer conferencing. Pragmatics and beyond New Series, 29-46.

Zappavigna, M. (2012). Discourse of Twitter and social media: How we use language to create affiliation on the web (Vol. 6). A&C Black.

Zappavigna, M. (2015). Searchable talk: the linguistic functions of hashtags. Social Semiotics, 25(3), 274-291.

Zappavigna, M. (2018). Searchable talk: Hashtags and social media metadiscourse. Bloomsbury Publishing.

Zhang, W., & Watts, S. (2008). Online communities as communities of practice: a case study. Journal of Knowledge Management.