8-8-2023

Variation Of English Intensifiers With Formality On The Internet

Joshua Baumgarten
Syracuse University

Abstract

Intensifiers are a class of adverbs that are noted to change relatively rapidly when compared to other parts of English. Frequently, their role as intensifiers results from the process of grammaticalization in which they take on this emphatic function. Over time, certain intensifiers have adopted indexicality, often carrying connotations of informality and femininity. *Very*, on the other hand, stands out from the other most frequent intensifiers (*really*, *pretty*, and *so*) due to its classification as a formal intensifier, evidenced by its favored usage in formal speech and written text. In the present study, the perceived formality or informality of these common intensifiers was examined with a corpus of Internet language, a relatively new register that is unique in the fact that the language found within is written, but generally informal. By examining the distribution of *very*, *really*, *pretty*, and *so* across differing levels of formality on the Internet, it was determined that *really* is the overall favorite intensifier, but that *very* maintains its position as the go-to intensifier in more formal situations. In addition, *pretty* and *so* were found to behave differently than the more common *very* and *really*, indicating that they are following a different path of grammaticalization.

Variation of English Intensifiers with Formality on the Internet

by

Joshua Baumgarten

(B.A., University of Pittsburgh, 2019)

Thesis

Submitted in partial fulfillment of the requirements for the degree of Master of Arts in Linguistic Studies

Syracuse University

June 2023

Acknowledgements

Many thanks to Dr. Corrine Occhino for your continued guidance, advice, and support. Thank you to Fei Wu, Dominic Grasso, and Dr. Calle Börstell for taking the time to assist me with Python and R scripts, and thanks to Dr. Rania Habib and Dr. Kenji Oda for their participation on my committee and their feedback. This would not be possible without the help of all of you.

Contents

Table of Figures

1. Introduction

1.1 Defining and studying intensifiers

In English, adverbs constitute perhaps the most diverse grammatical category. This is due to the large variety of forms and functions that appear in different contexts. The most relevant type of adverbs for the current study are intensifiers. Sometimes referred to as amplifiers (Quirk et al. 1985:567), intensifiers are found to belong to the category of adverbs of degree, which Biber et al. (1988) define as adverbs that "describe the extent to which a characteristic holds," and "can be used to mark that the extent of degree is either greater or less than usual or than that of something else in the neighboring discourse". In this explanation, intensifiers are specifically defined as "degree adverbs that increase intensity (551). Ito & Tagliamonte (2003) refer to intensifiers as "adverbs that maximize or boost meaning" (258). While various definitions exist in literature regarding adverbs, the dominant purpose of an intensifier is to increase, and often emphasize, the extent to which an attribution is understood to describe something.

In an examination of language change, adverbials – particularly intensifiers – can be an insightful focus of study. Intensifiers are said to be among the most rapid areas that undergo semantic change (Quirk et al. 1985:590; Peters 1994:269; Stoffel 1901). Peters (1994) cites "a taste for hyperbolic expression in language" as a cause for intensifiers' disposition to change, in that speakers use them to "be original, to demonstrate their verbal skills, and to capture the attention of their audience" (271). Explained to be creative discourse markers as well as indicators of group membership, he concludes that intensifiers eventually lose their group-marking function as they spread to other speech communities, resulting in rapid meaning change over time.

In addition, intensifiers appear to carry sociolinguistic meaning. Seen as both “vulgar” (Fries 1940) and feminine (Jespersen 1922) aspects of language, the indexicality of intensifiers may reveal a speaker’s intended level of formality (Tagliamonte 2016), age (Ito & Tagliamonte 2003), or stance towards a particular topic.

1.2 Intensifiers and grammaticalization

Frequently, the language change process that is seen in the case of intensifiers – and adverbs in general – is grammaticalization. This term, which has also been called grammaticization, was first defined by Meillet (1912) as “the attribution of grammatical character to an erstwhile autonomous word” (131). Meillet’s coinage was one of many early attempts to account for the formation of grammatical forms in a language. While there have been numerous interpretations of this process, grammaticalization is generally considered to entail a process during which a certain lexical form shifts away from its clear lexical meaning and adopts a more functional or grammatical one.

In English, adverbs appear to be a relatively common domain for grammaticalization to occur, and have been noted to undergo this process in a number of ways. Several *-ly* adjectives, for example, have undergone a semantic shift from their original descriptive meaning and towards an intensifying or emphatic one. *Strangely* emerged during the fourteenth century carrying a meaning that described unease, oddities, and of course, strangeness. By the seventeenth century, however, it had adopted a role as an intensifier when modifying an adjective:

(1) How **strangely** kynd are you ... I am **strangely** ioyed in the hopes you give us (1660s, letter, Thimelby; cited in Lewis 2020:6).

In the example above, the use of *strangely* is not remarking on the characteristic or quality of the adjectives "kynd" and "ioyed," but is rather increasing the extent to which these attributes describe "you" and "I." Lewis (2020) also provides the example of *luckily* which followed a distinct path of grammaticalization from *strangely*. A coinage dating back to the fifteenth century, *luckily* was an adverbialized form of the adjective *lucky*, and was used to evaluate an event, describing good fortune as a result of chance. The semantic scope of this adverb has expanded to include meanings such as "successfully" and even "fortunately". Unlike *strangely*, whose course of grammaticalization led to it frequently modify adjectives in an AdvAdj position, *luckily*'s direction through grammaticalization has resulted in its use as a marker of speaker's stance or attitude towards an event, generally occurring at the end of a statement in a VPAdv position:

(2) so I didn't lose my deposit **luckily**

(BNC2014, SY2Z; cited in Lewis 2020).

In this instance, *luckily* is seen to have been grammaticalized to adopt a stance-taking function, indicating a positive attitude towards an event. While the trajectories of *strangely* and *luckily* are quite different, they both entail a shift from an original lexical meaning to one that is more functional in nature. It is important to note that grammaticalization does not necessarily result in the loss of the original lexical meaning. Both *strangely* and *luckily*, for instance, can be used in their original senses to describe an action without intensifying it or attributing a speaker's attitude. Instead, grammaticalization can be thought of as an expansion of the scope in which a form may appear. It may be the case that the original meaning loses prominence in everyday speech, but Hopper & Traugott (2003) stress the importance in acknowledging grammaticalization as a shift, rather than sudden loss, of meaning (94-6).

This perspective is important to have when examining ongoing grammaticalization. Hopper (1991) proposes five principles that pertain to grammaticalization: layering, divergence, specialization, persistence, and de-categorialization. The first two particularly attest to the ways in which the original meaning of a grammaticalized form may shift. Layering refers to the fact that old forms within a functional domain in which a grammaticalized item emerges do not simply disappear as new ones appear, but rather they “may remain to coexist with and interact with the newer layers” (22). He exemplifies this principle with the English past tense. Originally involving a vowel alternation within a verb to describe the action as having occurred in the past (e.g., *drive/drove, take/took),* the addition of a [t] or [d] suffix was a past marker that emerged later, likely due to the grammaticalization of the verb *do* (Lühr, 1984). Despite the emergence of the more recent past marker, the older vowel change marker has continued to appear along with the suffix form across centuries, constituting coexisting functional “layers” (22-4).

The second of Hopper’s (1991) principles that attests to the shift of the original meaning of a word that has undergone is called divergence, and in some cases, can be thought of as a special form of layering. In instances of divergence, “the original form [of a grammaticalized lexical item] may remain as an autonomous lexical element and undergo the same changes as any other lexical items” (24). Divergence may constitute instances in which a single lexical item undergoes this process in one context, but does not undergo any change in another. To illustrate this, Hopper provides the example of Latin *habere*, which later became a future tense marker in French – *je chant**erai*** ‘I shall sing’ – as well as a lexical verb *avoir* ‘to have.’ At this stage, two layers of the original *habere* are visible: a future tense marker, and a lexical possession 'have.' Divergence takes place among the latter form *avoir*, which Hopper explains took on the role of an auxiliary to make the perfect aspect, yielding a sentence such as *j'ai chanté* 'I have sung.' Whereas the lexical verb *avoir* remains functional, in perfective contexts, its meaning has diverged to the auxiliary, as in the case of *j'ai*.

Thus, the presence of both layering and divergence can exhibit the footprint of grammaticalization on Modern French. Layering took place when the Latin *habere* grammaticalized into two separate, but concurrent, meanings in French. Divergence then took place when the second of these two meanings, the lexical verb *avoir*, maintained its possessive 'have' meaning in some contexts, but assumed an aspectual perfective 'have' meaning in others.

Another clear example of a grammaticalized form that exhibits layering and divergence is the Old English intensifier *swīþe*, meaning 'very, much, exceedingly,' whose trajectory is outlined by Méndez-Naya (2003) in her examination of its use in the Helsinki Corpus. She explains that this word is derived from the adjective *swīþ*, which meant 'strong, powerful,' forming an adverb with the meaning 'strongly, powerfully, violently' (378). This form later took on a role of intensification and emphasis, as is shown in the sentence:

(5) forðon ðe ic eom ***swīþe*** mildheort.
because I am very merciful
(QO2_STA_LAW_ALFLAWIN: 38; cited in Méndez-Naya 2003:380)

Despite its emergence as an intensifier, *swīþe* developed an additional meaning of 'quickly, fast, soon', which was used as a manner adverb, rather than a degree adverb. It is explained that this additional meaning was context-dependent at first, but "soon became inherent to the adverb" (383), as is shown in the following example:

(6) Do *swīþe* sei me for ich chulle lowse þe & leten when me þuncheð.
Do quickly tell me because I will free you & let when-to-me seems.
(QM1_NN_BIL_JULME: 102; cited in Méndez-Naya 2003:383)

Méndez-Naya explains that the original intensifier function of *swīþe* is not observed after around 1250. We can see *swīþe*, once the most common intensifier in English, exhibit both layering and divergence. Evidence of layering is shown by the emergence of other adverbs such as *ful*, *wel*, and *rihte*, which came to be during the lifetime of *swīþe*. Méndez-Naya suggests that these newer intensifiers had "more emotive force" used to replace *swīþe* which "had lost its former expressivity" (387). That is to say, newer forms continuously emerged to eventually replace older grammaticalized forms, as is the case in the demise of *swīþe*.

Divergence, on the other hand, is observed in the development of the 'quickly' interpretation of *swīþe*, whose trajectory was distinct from its 'very' meaning. Whereas the intensifier or degree function faded from use, its manner adverb use became the primary interpretation of the word. Acknowledging layering and divergence as evidence for a grammaticalized form allows us to identify instances in which grammaticalization may have occurred or be occurring. Furthermore, because the grammaticalized meanings can be derived from the original meanings of words, these new meanings must not be arbitrary, and can therefore also be helpful in identifying cases of grammaticalization (Hopper & Traugott 2003: 94).

It is evident that the intensifier meaning of *swīþe* is semantically connected to the original lexical meaning of ‘strong’ that it carried, but that the original meaning has faded. This is not uncommon with intensifiers, and can be seen in the more modern example of *very*, whose original meaning of ‘true, real’ took on the role of emphasis as it transitioned from this lexical meaning to one of intensification. Later, its scope of use expanded to attributive adjectives before fully taking on its intensifier role (Ito & Tagliamonte 2003; see also Mustanoja 1960). Over time, the use of *very* became constrained to this intensifier function, exhibiting Hopper’s principle of specialization and revealing *very*’s significant progress within the process of grammaticalization.

1.3 Noted changes in intensifier usage

Several examples have been noted regarding changes in how intensifiers are used across time (Ito & Tagliamonte 2003; Tagliamonte & Roberts 2005), as well as across register (Tagliamonte 2016). In their study of York English, Ito & Tagliamonte (2003) found that while *very* was the most common intensifier used, it was primarily used by older speakers. On the other hand, *really* was shown to be rapidly gaining popularity, but almost exclusively by the younger speakers. The authors explain that both *very* and *really* serving as intensifiers is the result of grammaticalization, albeit at different points in this process. *Very* is explained to be at a more advanced point of grammaticalization than *really*, evidenced by its narrower context of use (having a stronger preference to modify predicative rather than attributive adjectives) as well as the fact that it was not found to maintain its original lexical meaning of ‘truly’, whereas *really* was found to do so.

In examining the representation of everyday spoken English on television, Tagliamonte & Roberts (2005) looked at intensifier use across eight seasons of the American sitcom *Friends*.

They found many similarities between intensifier use in the show and that of contemporary spoken English. Notably, *really, very,* and *so* were the three most frequent intensifiers. Additionally, the rates at which these intensifiers occurred across the show's runtime indicated that *very* was decreasing in use, being replaced by *really,* and ultimately *so,* which was noted to be most common among younger generations. These findings corroborated other trends noted in intensifier studies, including *so*'s rise to prominence among young speakers in England (Ito & Tagliamonte 2003) and Canada (Tagliamonte 2004, cited in Tagliamonte & Roberts 2005).

The *Friends* findings also have implications from a sociolinguistic standpoint. Tagliamonte & Roberts note that the incoming intensifiers *really* and *so* were heavily favored by female characters, whereas the distribution of the more dated *very* was equal across gender. They cite both Labov's (1990) Principle II of language change as well as female speakers' tendency to use "more emotional language than men" as possible explanations not just for female characters' greater use of intensifiers in general, but their preference for the nonstandard incoming forms (288-90).

1.4 The indexicality of intensifiers

Beyond a connotation of feminine speech, intensifiers are commonly considered to be informal aspects of language. Fries (1940) divided English intensifiers into "standard" and "vulgar" categories. The majority were considered "vulgar", however *very* was deemed to be "standard". Other accounts of intensifiers' formality have noted *very* as being an outlier when compared to the nonstandard majority. Tagliamonte (2016) examined intensifier use across different registers that varied in medium as well as formality. Using the Toronto Internet Corpus (TIC), as well as a corpus of University of Toronto students' written work, Tagliamonte looked at the usage of intensifiers in three internet registers – increasing in formality: SMS text messages, instant messaging (IM), and email – as well as formal writing.

In the TIC, there was a clear connection between *very* and formality, whereas *really* and *so* showed a tendency to appear in more informal contexts. *Very* was the only intensifier to appear in the formal written register, whereas among the four most common English intensifiers (*really, so, very, pretty*), it was by far the least common in the three internet registers. On the other hand, the so-called "vulgar" intensifiers occurred in each of the informal registers, with *so* being the most common. Notably, as the formality decreased across these registers, with SMS being the most informal, the frequency of the incoming *so* increased.

Besides the degree modification meanings held by intensifiers, there are clear social connotations associated with its use, as evidenced by various contributors to their discussion. As early as the turn of the twentieth century, numerous authors had commented on female speakers' fondness for the use of intensifiers (Stoffel 1901, Jespersen 1922). Jespersen (1922) remarks that "the fondness of women for hyperbole will very often lead the fashion with regard to adverbs of intensity, and these are very often used with disregard of their proper meaning" (250). This comment is rich not only in social evaluation towards women, but opinions regarding intensifier use in general as well as the employment of a grammaticalized form. Jespersen explicitly associates the employment of these "adverbs of intensity" with hyperbolic, expressive speech, and explains that female speakers' more prominent usage of speech of this nature predisposes them to use these intensifiers.

Furthermore, without explicitly recognizing the process of grammaticalization, Jespersen appears to negatively evaluate the outcome of it. Citing phrases such as "awfully pretty" and "terribly nice," he claims the female usage of *awfully* and *terribly* to be a "disregard of their proper meaning." Here, he is remarking on the fact that these two adverbs are no longer being used in their original lexical contexts, and he acknowledges that their meanings as intensifiers no longer equal the sum of their parts. That is, *awfully* no longer implies that an action was done in a very bad manner, rather it has become an intensifier that emphasizes the degree to which something is "pretty." In this new sense, *awfully* does not simply mean *awful + ly*. In his other example, *terribly* has undergone a similar trajectory, now functioning the same way. Rather than noticing that the semantic scope of these adverbs has simply expanded to include a functional, stance-marking role, he associates this change with a woman's desire for hyperbole and denounces it.

Although these commentaries are over a century old, similar negative attitudes towards grammaticalized forms are still quite common today. For example, a submission on the Reddit page dedicated to unpopular opinions from December 2022 is titled "I don't think definitions of words should change just because people are commonly using the word wrong." The post focuses on the word *literally* and denounces its emphatic use. The poster acknowledges that this change in meaning is a result of speakers using the word in this new "wrong" sense, but does not recognize that this is a perfectly fine process that takes place in language. Instead, he criticizes the new meaning associated with the word and chalks it up to speakers' errors rather than noting that speakers' usage of the word determines how its meaning may expand. This type of opinion has continued to be present since Jespersen expressed it a hundred years ago, and is still quite common. In fact, posts similar to the one from December 2022 are plentiful on Reddit, and the post in question has since been deleted for not being unpopular enough.

Continuing his discussion on the female tendency for intensification, Jespersen cites *so* as an "intensive which has also something of the eternally feminine of it" (250). Stoffel (1901) had previously made the connection between women and the use of *so* as an intensifier, citing not only that "ladies are notoriously fond of hyperbole," but explaining that the phonetic quality of the word better aligns itself with a female speaker's desire to be expressive:

> "a strong-stressed *so*, with the first element of the diphthong in it abnormally long, before an adjective, from a lady's lips, conveys a sense widely different from a strong-stressed *very* under the same circumstances. Compare, for example, the almost passionate force of «You are so kind!» with the comparatively tame and colourless «You are *very* kind!»" (101).

Stoffel's explanation offers a glimpse into his period's social evaluation of intensifiers. Not only does he attribute *so* as being an intensifier that is associated with feminine speech, but he also remarks on the fact that this expressive or "strong intensive" (101) function of the intensifier is a recently emerging form:

> "[t]his exceptionally strong-stressed *so*… is a special feature also of the female epistolary style of our time, but it is difficult to find examples of it in literature before the present century. In contemporary English, however, it is very frequent" (101-2).

He, like Jespersen twenty years later, acknowledges the emergence of an intensifying function of an already existing adverb. Both authors also attribute this specific usage of adverbs to women's desire for hyperbole. Stoffel continues further, associating *so* with children and "ladies men" (102). That is, the intensifying function of adverbs seems to be all but masculine in the minds of these authors. Furthermore, his own use of *very*, as well as his description of a *so* expression being an "almost passionate force" whereas a *very* expression is "tame and colourless" in comparison suggests that Stoffel judges the level of formality associated with certain intensifying forms. His descriptions indicate that he considers *very* as a more formal choice when compared to *so*, which in his eyes, is informal, feminine, hyperbolic, and "employed… especially in colloquial usage" (101).

Judgements, like Stoffel's, regarding the (in)formality of intensifiers are fairly common. As noted above, Fries (1940) has made similar remarks, categorizing *so* as well as other common intensifiers (i.e., *really* and *pretty*) as "vulgar" compared to the "standard" *very*. Literature of this topic tends to show agreement in the consideration of *very* as more formal when compared to other intensifiers (Stoffel 1901, Fries 1940, Tagliamonte 2016).

These discussions on the usage of intensifiers are quite significant, as they reveal that this particular class of adverbials is socially salient. That is to say, the emergence of an intensifier function of certain adverbs is not only recognized by speakers, but is also socially evaluated. Certain intensifier forms are shown to index social meanings, namely femininity and colloquial or formal speech. This indexicality is what has led to the focus of this current study. In this section, I have reviewed not only how intensification is seen as a sociolinguistic phenomenon, but how intensifiers and adverbs in general are prone to change over time. As such, I am interested in pursuing this topic, with my research guided by the following questions:

1. What intensifiers are most commonly used today? Are these consistent with trends found in previous studies regarding which are rising or falling in use?
2. How are these intensifiers being used? Are there any noted changes in their function, such as the addition of a stance-taking or speaker-attitude-revealing role?

3. How does the topic and formality (or lack thereof) influence the use of intensification and the choice of intensifier used?

The findings from the studies above reveal several factors to be considered when studying how intensifiers vary in usage: (1) while once a popular choice, *very* is declining in use and rapidly being replaced with *really* and *so*, the latter being the most common intensifier in recent studies; (2) adverbs appear to pick up their intensifying function through grammaticalization, and the distribution of different individual intensifiers in English may reveal their progress within this process; (3) there are evident distinctions in the formality associated with different intensifiers, indicating that speakers are aware of the indexicality of intensifiers when it comes to formal and informal speech and writing; and (4) female speakers are found to use informal intensifiers *really* and especially *so* at a higher rate than male speakers, which may be explained by women's more "emotional" speech or Labov's Principle II of language change. These trends serve as the foundation for hypotheses in the present study.

1.5 The Internet as an informal written register

Considering their significant social salience with regard to different levels of formality, an examination of adverbs warrants consideration of the formality of the register in which they occur. The internet in particular provides a unique and relatively new register in which the usage of intensifiers can be studied. Historically, written language has tended to be more formal than spoken language. With the rise of the internet over the last few decades, a new informal written register has emerged as a source of an immense amount of recorded language. The medium of communication, often involving small or incomplete keyboards, as well as the facilitation of rapid conversation on the internet encourage a more casual or informal manner of writing.

In order to investigate how intensifiers are used within this medium, Reddit, a forum-based website centered around conversation surrounding thousands of specific topics, was chosen as a source for the written language to be examined in the current study. Reddit contains a highly organized and extensive conglomeration of examples of this informal written language, therefore providing ample opportunities to encounter intensifiers. Based on previous findings of intensifier usage, the predictions for this study included higher use of *really* and *so* when compared to other intensifiers, but a preference for *very* when conversation is more formal.

2. Methodology

2.1 Approaching COCA and COHA

In order to look into the frequencies of common English intensifiers, a corpus of data from Reddit was utilized. However, before examining Reddit, frequency tests for the four most common intensifiers – *very*, *really*, *so*, and *pretty* – were first performed in the Corpus of Contemporary American English (COCA) and the Corpus of Historical American English (COHA). COCA data was examined in order to establish a baseline understanding of the overall distribution of these intensifiers. Tests were conducted by searching for the relevant tokens specifically in intensifier positions.[1] In order to narrow down this structure, instances of these intensifiers followed by adjectives were noted, following the format *intensifier* + ADJ. For each *intensifier* + ADJ construction searched, the overall amounts of tokens were recorded, and the most frequent adjectives used in these constructions were noted with their corresponding token counts as well.

[1] The frequency tests performed were modeled after those done by Kim & Moon (2014) in their examination of SKT constructions.

Frequency tests were also performed in COHA in order to visualize trends in the usage of these intensifiers as well as to be able to compare the trends with previous findings regarding intensifiers' changing uses. Once again, the overall amount of tokens for each *intensifier* + ADJ construction were noted. COHA provides a breakdown of token count by decade back to 1820, allowing for a diachronic view of the frequencies in question. In addition to noting the token counts of the most common adjectives that are modified by the intensifiers, the year in which the use of each adjective peaked as well as the token count for the peak year were recorded. The results from the tests run in COCA and COHA were not intended to be compared to those from the tests of Reddit. Instead, these preliminary analyses were to visualize and understand how the intensifiers of study tend to appear in various contexts in English.

2.2 Reddit data collection and organization

With the results of the COCA and COHA frequency tests providing a baseline perspective on recorded instances of intensification, a corpus of text from Reddit was examined in order to study the usage of these intensifiers on that website. There are two main reasons why Reddit was chosen as a focus of study. First, it serves as a very large collection of informal written language. As intensifiers are frequently considered to be informal or vulgar aspects of spoken language, Reddit provides an opportunity to evaluate if a setting of informal language promotes their usage. As of 2021, Reddit boasted 52 million daily users, with 47.82% being from the United States. In fact, the four most represented countries on Reddit are predominantly English-speaking, with the US strongly leading. Therefore, there is no dearth of American English on the website. Additional demographic statistics include a majority male and young userbase, with 62% of users being male, and 36% being between 18-29 years old (as of 2021). 22% of users are 30-49 years old, and only 13% are 50 or older (Dean 2023).

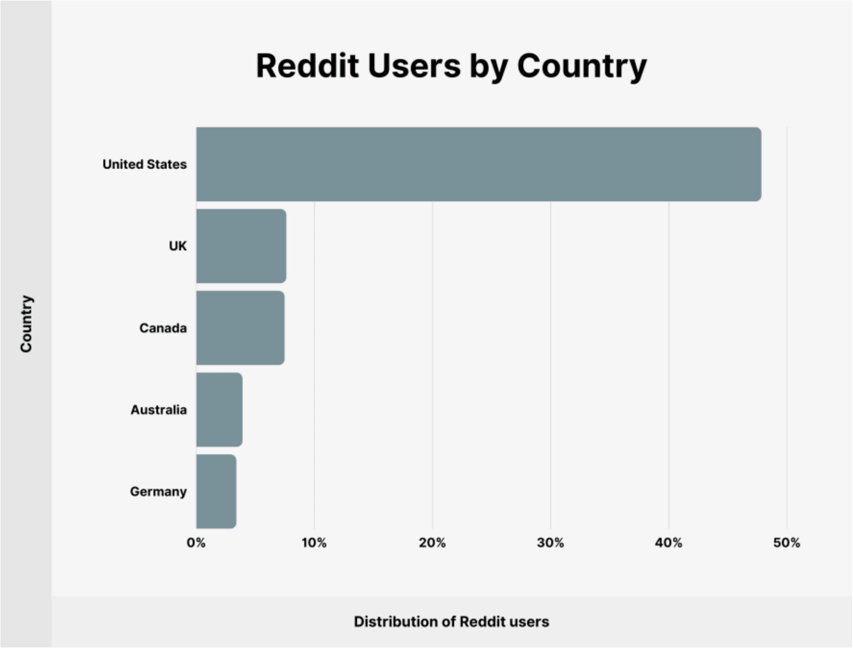


FIGURE 2.1: Distribution of Reddit users by country. Source: Statista

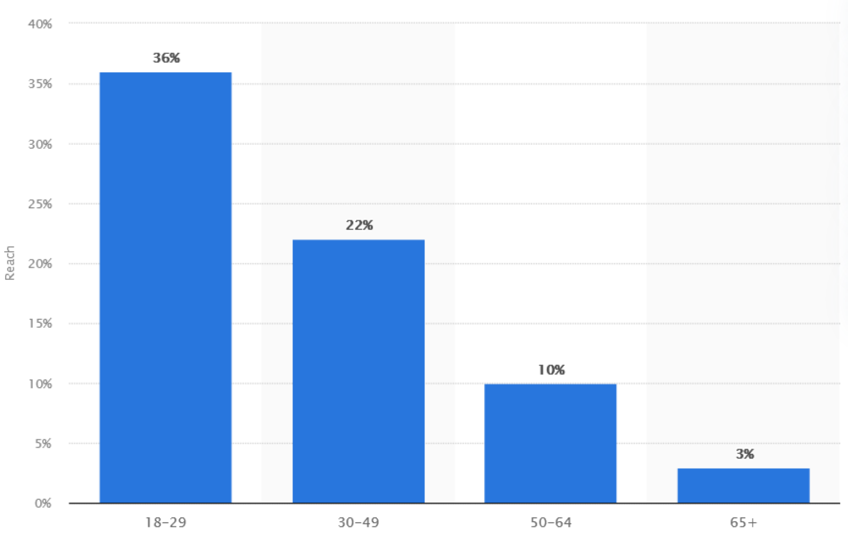


FIGURE 2.2: Age demographics for Reddit. Source: Statista

The next reason Reddit was chosen is the variety of topics to which users gather that are found across the website. As of 2021, the site had around 3,125,000 subreddits – individual forums dedicated to a specific topic. These subreddits and the website as a whole contain a great variety of conversational topics and styles, revealing a sort of spectrum of formality, in which certain pages encouraging more formal or informal conversation can be identified. According to a 2019 survey of 2,100 Reddit users, 73% cited entertainment as their reason for visiting the website, and 43% cited news as one of their motivations (Dean 2023). These two most frequent reasons for using Reddit illustrate well how both informal and formal conversations may occur on Reddit. While the internet as a whole may be considered an informal register, these further divisions of (in)formality on Reddit in the form of subreddits may provide more detailed insight into how intensifier usage may vary within this register. Figure 2.2 illustrates how Reddit forums are structured.

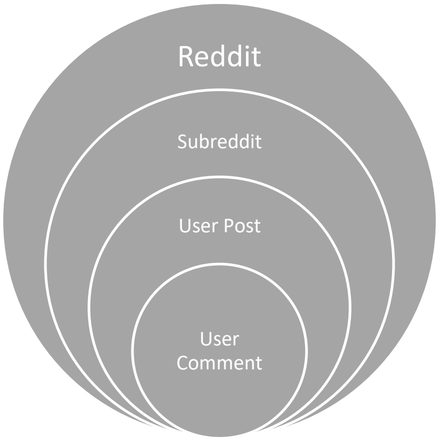


FIGURE 2.3: The embedded structure of Reddit. The overall website is made up of subreddits – individual fora dedicated to a specific topic. Within each subreddit, users can create posts, and other users can comment on these posts.

For the current study, classifications of subreddits based on expected formality of conversation within were established. It is important to note that this classification is a fairly subjective practice, as one cannot always reliably find subreddits that explicitly state whether text will be formal or informal. Observations were made about various subreddits allowing for a general categorization of subreddits into a more Formal group, an explicitly Informal group, as well as a Neutral group in which the intended formality of the conversation was not explicit, but whose topic of focus was more informal in nature. Several factors went into this categorization, including the topics of discussion, the ways in which discussion was structured, as well as the attitudes posters had towards the subreddit itself. The rules of different subreddits were also useful in determining the page's expected formality. Several pages, such as r/AskDocs and r/AskAnthropology, followed the structure of a poster asking a question and commenters providing answers. Explanation-based subreddits such as these frequently prohibit commenters from responding in joking or unserious manners. This rule, in addition to the explanatory nature of the page, appeared to encourage conversations that would be more formal. As such, these subreddits were grouped as Formal ones.

The Neutral and Informal subreddit categories were more difficult to classify, with distinctions between their formalities being more nebulous. Informal subreddits were considered those that outwardly indicated the informal nature of the content that would be present on the page. A clear example of this is r/CasualConversation, whose name overtly reveals that the language is expected to be informal, or casual. Other pages that may not have an indicator such as "casual" in their names also revealed themselves as informal communities, either due to a vulgar topic or an understanding that the conversation is not meant to be taken seriously. r/shitposting is a crude example of this type of subreddit, a page that centers around stupid jokes and comment chains, and whose first rules listed are “take it easy” and “don’t be a c***”. While the topics of subreddits such as r/CasualConversation and r/shitposting are both understood to be informal, the attitudes participants had towards engagement on these pages differ. r/CasualConversation users intend to actually engage in conversation with others, whereas users on r/shitposting do not take the page seriously, but rather participate in an almost performative nature that reinforces the silliness found within this subreddit. These observations of how different Informal subreddits encouraged different types of participation led to the decision to split the Informal category in two. Informal A became a category of subreddits consisting of explicitly informal pages with users truly attempting to develop conversation. Informal B contained explicitly informal pages in which the pages’ users treat them as a farce, being purposefully silly or foolish.

To establish a more central classification, subreddits that did not have any indication of being formal or informal were placed into a Neutral category. These, by nature, were on the informal side of this spectrum, mainly due to the topic of conversation and the nature of the website itself. Specific topics such as cast iron cookware, baseball, videogames, music, etc. do not encourage formal conversations or explanations, while also not acknowledging that any ensuing conversation is not meant to be taken seriously. These types of pages do have rules, however they do not prohibit jokes in the comments. r/NFL, for example, does prohibit original content posters from creating posts that are jokes, but funny or joking comments left on existing posts are allowed, and appear to be fairly common. r/AskReddit, a particularly popular subreddit, would also be eligible for this group. It is worth mentioning that there are several pages following the “AskX” naming paradigm. While more formal pages such as r/AskScience or r/AskHistorians follow this pattern, this naming scheme does not dictate formality or explanation-based discussion. Instead, they focus on a question-response structure, with individual subreddits determining how formal or informal they would like the conversation to be. Table 2.1 shows the spectrum of formality along which different subreddits were classified, as well as the criteria behind this classification. Table 2.2 shows the different categories of formality used, as well as the subreddits analyzed within each category.

Level of Formality	**Subreddit Category**	**Criteria**
↑ Formality ↓	Formal	- Academic, scholarly, or professional topics - Stricter rules on who can post and what types of posts are permitted (e.g., no jokes)
	Neutral	- Centered around a particular topic, but not explicitly academic or explanatory in nature - More liberal rules for posting; jokes permitted
	Informal A	- Explicitly casual in nature, often having the word "casual" in the name - Comments are not restricted to formal responses (e.g., jokes are permitted) - Users intend to discuss a topic in good faith
	Informal B	- Very casual in nature - Few rules restricting the way users can contribute - Users tend not to attempt to have good-faith conversation, rather will joke, parody, and act explicitly unserious in an almost performative way

TABLE 2.1: The spectrum of formality used to classify different subreddits from Formal to Informal B. Criteria used to make these classifications are noted.

Categories	**FORMAL**	**NEUTRAL**	**INFORMAL A**	**INFORMAL B**
Subreddits	AskDocs AskEconomics history AskAnthropology AskBiology AskSocialScience linguistics Physics AskVet	castiron eagles CozyPlaces FoodPorn AnimalsBeingDerps Spiderman calvinandhobbes nevertellmetheodds	CasualFilm CasualAskreddit casualworldnews casualnintendo casualknitting casualconversations CasualConversation	shitposting Gamingcirclejerk okbuddyretard metacirclejerk surrealmemes circlejerk

TABLE 2.2: The categories of subreddits analyzed based on formality, and the subreddits constituting said categories.

In order to collect data from Reddit, the Reddit corpus from Cornell's ConvoKit website was utilized through scripts written in Python coding language. This corpus contains the content of 948,169 subreddits spanning a time period from an individual subreddit's creation until October 2018. The information available within this corpus is substantial and varied, including every comment posted, the username of the comment's author, the time at which a comment was submitted, and much more. For the current study, the focus was simply on the comment, or utterance, itself.

A script was written in Python that accessed the corpus and looped through every utterance within a specified subreddit. This program would then append all the utterances from the selected subreddits that contain an intensifier to a data frame to be exported to a Microsoft Excel spreadsheet for review. Additionally, the program would also output the amount of tokens of *very*, *really*, *pretty*, and *so* found within the subreddits.[2]

[2] The immense size of some of the subreddits found within the corpus meant that hardware limitations played a role in the data collection. The largest subreddits, those being the most popular on all of Reddit, occasionally had too much data for the program to properly output to an Excel spreadsheet. As a result, subreddit selection had to take this into account.

Subreddits were chosen in a way that encouraged diversity in conversation topic as well as a large sample size of utterances. Within each category of formality, subreddits were selected to be run through the Python program in accordance with the specifications of what constituted a Formal, Neutral, or Informal page. To obtain a reasonable sample size, a goal of five million utterances per category was sought. This goal was established based on the size of r/CasualConversation, which served as a prototypical subreddit within the Informal A category. Its explicitly informal nature as well as the prevalence of casual, normal conversation on any topic meant that this particular subreddit was a prime example of the type of language that can be promoted within an internet register. As this one subreddit contained 5,879,435 utterances, it was important to reach a similar sample size in other categories as well. Certain subreddits such as r/CasualConversation and r/history were quite large and met this five million utterance goal by themselves. In cases where this utterance goal was reached by just one subreddit, additional smaller subreddits within that same formality category were also analyzed in order to diversify the types of discussions in which utterances would be found.

Using multiple subreddits allowed for a greater variety of conversation topics and an overall greater diversity of users and commenters. This meant that patterns of intensifier use from one subreddit would not be assumed to constitute the typical style of commenting for any Formal, Neutral, or Informal subreddit. Instead, the patterns found in one subreddit could be compared to those from another within the same category, painting a clearer picture of the overall intensifier use within a specific register of formality. It is important to note that variation of formality within a subreddit is possible; that is, a subreddit categorized as Formal may certainly have some informal conversation, and vice versa. However, Reddit users typically participate in a particular subreddit with the goal of engaging with its focus and following the guidelines and expectations set by that subreddit. Overall, an individual subreddit will have a main, predominant form of engagement. For the purposes of this study, therefore, the nature of the interactions within a subreddit is assumed to conform to the level of formality assigned to that forum.

Once a sufficient amount of subreddits were run through the program, another spreadsheet was created to organize all the outputted data. The spreadsheet was organized by formality, and within each category, the amount of utterances and tokens of the intensifiers studied were noted. The overall token counts across each entire category were calculated as well.

3. Results and Analysis

3.1 Results from COCA and COHA frequency tests

COCA is a corpus made up of over one billion words from a variety of sources ranging from 1990 to 2019. Per the corpus's website, the data comes from television and movie subtitles, transcripts of spoken unscripted conversation, fictional works including short stories, novels, and movie scripts, magazines, newspapers, academic journals, blogs, and other web pages. Therefore, the data from these preliminary frequency tests comes from many different registers and levels of formality, while still being relatively recent examples of spoken and written English. The results from the frequency test from COCA can be seen in Table 3.1.

Intensifier String	**Tokens**	**Adjectives**	**Tokens**
very + ADJ	688,083	good	42,979
		important	23,035
		different	18,100
		difficult	15,540
		nice	11,901
		hard	10,090
so + ADJ	474,461	good	18,055
		sorry	16,272

		bad	13,263
		hard	9,123
		happy	8,784
		important	7,901
really + ADJ	177,116	good	20,490
		bad	6,358
		hard	6,002
		nice	5,817
		important	5,610
		great	5,438
pretty + ADJ	130,427	good	22,802
		sure	8,669
		cool	3,089
		clear	2,792
		bad	2,718
		big	2,324

TABLE 3.1: Tokens of intensifiers in a pre-adjectival position as found in COCA, as well as the most common adjectives used with these intensifiers.

From the COCA test results, *very* (688,083 tokens) clearly appears to be the most frequently used intensifier, exceeding *so* (474,461 tokens) by over 200,000 tokens. On the other hand, *really* trails quite significantly, with 177,116 tokens, and *pretty* is the least frequent of the four, making up 130,427 tokens. As these results are from many different registers, they are not necessarily indicative of the tendency for a particular intensifier to be used in a certain level of formality, but rather the overall popularity of that intensifier. Therefore, we can see that *very* is the most popular choice across all the registers that make up COCA, with *really* and *pretty* occurring far less frequently.

While COCA does not allow for strings such as *intensifier* + ADJ to be compared between the different registers found in the corpus, this type of comparison is quite easy for individual words. The prominence of *very* was slightly surprising given the noted decrease in its usage in previous studies as well as the relative recency of the data in COCA. To further examine the use of *very*, its frequencies in two registers in COCA that are more formal – NEWSPAPER and ACADEMIC – were compared to registers that were expected to be more informal – WEB-GENL and BLOG. The same frequency comparisons were performed for *really* to test if certain registers favored one intensifier over another.

	Web Sources			**Formal Written Sources**		
Source	**WEB-GENL**	**BLOG**	**Total**	**ACADEMIC**	**NEWSPAPER**	**Total**
Total Words	124,253,679 words	128,613,294 words	252,866,973 words	119,790,456 words	121,741,989 words	241,532,445 words
***very* tokens**	133,099 tokens	148,231 tokens	281,330 tokens	66,171 tokens	77,003 tokens	143,174 tokens
	1,071.2 per million words	1,152.5 per million words	1,112.6 per million words	552.4 per million words	632.5 per million words	592.8 per million words
***really* tokens**	106,289 tokens	143,411 tokens	249,700 tokens	15,117 tokens	55,767 tokens	70,884 tokens
	855.4 per million words	1,115.1 per million words	987.5 per million words	126.2 per million words	458.1 per million words	293.5 per million words

TABLE 3.2: A comparison of tokens of *very* and *really* in web sources (general and blog) and formal written sources (academic and newspaper) from COCA.

It is important to note that comparing amounts of tokens of *very* versus *really* within the different types of sources is not entirely revealing of their use, as there is not an equal amount of words included in each COCA source. For example, there were 252,613,294 words within the WEB-GENL and BLOG sources when compared to the 241,532,442 from the ACADEMIC and NEWSPAPER sources. In order to ensure that a higher count of one intensifier over another is not simply due to a larger corpus size, the amount of tokens of a particular intensifier per million words – a figure provided by COCA – was used for comparison.

The findings from these tests show results that are slightly unexpected. Recall that my predictions for intensifier use in this study were a preference for *very* in formal conversation, but a higher use of *really* and *so* overall, as stated in Section 1.5. For these narrower COCA analyses, I also predicted a preference for *very* in the more formal sources, but a preference for *really* in the web sources. It was found that *very* was in fact the preferred intensifier in all four types of sources, also occurring more frequently in the web sources than the formal ones. This particular finding was surprising due to the frequent consideration of *very* as a formal intensifier and the assumed informality of Internet speech. In the web sources specifically, *really* did occur quite frequently at 987.5 tokens per million words, but still trailed *very,* which accounted for 1,112.6 tokens per million words.

In the formal sources, the preference for *very* as well as the relatively much lower use of *really* were less surprising. Compared to *very*'s 592.8 tokens per million words in the ACADEMIC and NEWSPAPER sources, *really* had just 293.5 tokens per million words, the lowest per-million-words figure from this study. While the preference for *very* over *really* in the web sources was not expected, the distinct frequencies for *very* and *really* in the formal sources indicate a strong preference for the former in this type of writing, a trend noted previously (Tagliamonte 2016). Finally, when looking at overall intensification in these two groups of sources, the more informal web sources show a higher propensity of the use of intensifiers than the formal written sources, which may support prior claims that the use of intensifiers is more characteristic of informal or colloquial language.

The adjectives that most frequently paired with each intensifier in COCA also allow for some interesting analysis. For each of the four intensifiers, *good* was the most common adjective that was emphasized. In most cases, the second most frequent adjective trailed *good* significantly, however *so* is an exception. The second most frequent adjective for *so* was *sorry* making up 16,272 tokens when compared to *so good* at 18,055. As *good* occurred most frequently with each intensifier, not much is revealed about its use. Rather, it is likely just the most common adjective in English, making its heavy usage with each intensifier quite probable. Therefore, the discussion of these results will not focus on how *good* is used.

With the high frequency of simple adjectives such as *bad*, *nice*, and *hard* in English, it is not surprising to see many occurrences of them with the four intensifiers. While these frequent adjectives, as well as *important*, occur with several of the intensifiers, we can see that their relative frequencies vary depending on which intensifier is used. For example, after excluding *good* (which as we saw occurs most frequently with each intensifier) *bad* is the most frequent adjective with *really*, the second most with *so*, and the fourth most with *pretty*. On the other hand, *important* is the fourth most frequent adjective with *really*, the fifth with *so*, and the most frequent with *very*. Although many adjectives are found to occur with multiple intensifiers, the variation in the orders of each intensifier's most popular adjectives may reveal a preference for one particular adjective to appear with a certain intensifier. For example, a speaker wanting to emphasize the adjective *hard* may be more likely to choose *really* over *so* or *very* when intensifying. Table 3.3 illustrates these patterns, showing how adjectives shared between multiple intensifiers may still vary in their relative frequencies with each individual intensifier.

Intensifier	***very***	***really***	***pretty***	***so***
Most frequently paired adjectives (in order of frequency)	important	bad	sure	sorry
	different	hard	cool	bad
	difficult	nice	clear	hard
	nice	important	bad	happy
	hard	great	big	important

TABLE 3.3: Most frequently paired adjectives (besides *good*) with each intensifier. Adjectives shared between multiple intensifiers are color-coded.

Another observable pattern regarding the adjectives that are commonly found with these intensifiers involves the complexity of said adjectives. The origin of words has been previously noted to affect how English speakers perceive a word's formality or complexity. Levin, Long, and Schaffer (1981) found that participants in formal settings would opt for the use of words originating from Latin rather than from Anglo-Saxon roots. Additionally, it was found that these participants linked formality with words that were less frequent. The authors summarize their findings, stating that "formality is defined by our subjects as Latinate words that are not frequent" (171). Within the present glimpse into intensifier-adjective pairing, an adjective may be thought of as more complex if it is polysyllabic and of a Latin root when compared to common, monosyllabic adjectives such as *good*, *bad*, and *hard*.

Interestingly, while *very* still occurs with these common adjectives, it shows a tendency to modify adjectives that are more complex when compared to the adjectives modified by *so*, *really*, and *pretty*. For example, *important* (23,035 tokens), *different* (18,100 tokens), and *difficult* (15,540 tokens) were all within the top five most frequent adjectives modified by *very*. While the data does have instances of these adjectives pairing with the other three intensifiers, the frequencies are significantly lower than they are with *very*, and are overshadowed by simpler, often monosyllabic adjectives. This pattern may be indicative of the nature of the language specific to certain registers found in COCA, or may perhaps reveal a preference held by speakers to use *very* when emphasizing or intensifying an adjective that is considered to be more complex or formal. The pattern noted above appears to be consistent with the findings of Levin, Long, and Schaffer. We can see that in addition to their lower frequency when compared to some of the adjectives used with the other intensifiers, many of the adjectives found to pair with *very* derive from Latin. While these findings are not conclusive, the preliminary results from the COCA frequency tests hint at a higher frequency of less common, Latin-derived adjectives when modified by *very*. Therefore, speakers may more strongly associate *very* with formal language than they do for *so*, *really*, or *pretty*.

Similar frequency tests were also performed in COHA in order to gain a more diachronic perspective of intensifier-adjective bigrams. The results from the frequency tests from COHA can be seen in Table 3.4. Because the most common adjective for each intensifier was *good*, a sixth intensifier-adjective pair was noted.

Intensifier String	**Tokens (top 100)**	**Most Frequent**		
		Adjective	**Total Tokens**	**Peak Yr./Count**
very + ADJ	172,007	**good**	16,134	1960 – 1,451
		different	6,457	1870 – 476
		important	5,396	1960 – 503
		nice	5,279	1960 – 657
		small	4,673	1880 – 288
		large	4,613	1880 – 331
so + ADJ	133,727	**good**	7,477	2010 – 684
		great	6,879	1880 – 530
		bad	5,399	2010 – 517
		happy	4,106	1930 – 300
		long	4,087	1940 – 305
		glad	4,044	1880 – 344
pretty + ADJ	25,071	**good**	6,451	2000 – 733
		sure	1,948	2010 – 447
		little	1,560	1870 – 132
		bad	1,064	1950 – 104
		young	609	2000 – 58
		big	574	2010 – 79
really + ADJ	18,533	**good**	2,475	2010 – 686
		sorry	945	2010 – 191
		great	817	2000 – 148
		nice	786	2010 – 217
		bad	708	2010 – 181
		important	697	2010 – 139

TABLE 3.4: Tokens of intensifiers in a pre-adjectival position as found in COHA, as well as the most common adjectives used with these intensifiers and the peak year for each adjective's use.

The data from COHA reveals a significant discrepancy between the use of *really* as an intensifier and *very* and *so*. A search for the *really* + ADJ construction returned only 18,533 tokens, whereas *very* and *so* returned 172,007 and 133,727 tokens respectively. While the degree to which *really* trails the use of *very* is significant, the fact that it is less common in the corpus is not entirely surprising. Considering the fact that the *very* tends to be more formal, its prevalence in a written register dating back 200 years is not unexpected. Interestingly, even *pretty* occurs as an intensifier more frequently than *really* in this corpus.

For each of the intensifiers tested, *good* was once again the most frequently modified adjective. Like the results from COCA, we can see similar patterns of lower-frequency, Latin-derived adjectives occurring more commonly with *very* than with other intensifiers. While *great*, *bad*, and *nice* were more frequently emphasized by *really* and *so*, more complex adjectives such as *important* and *different* were among the most common to follow *very*.

It should be noted that token count was not the only data examined from the COHA results. The benefit of this specific corpus is the diachronic perspective it offers. While token count is certainly a helpful metric to gauge overall usage of the adverbs in question, it is not necessarily helpful in revealing trends or changes in their usage. The time data provided by COHA provides insight into how the uses of these intensifiers have evolved. Notably, across the corpus, the peak years in which certain adjectives were modified by *very* tended to be much earlier than those in which adjectives were modified by *really*, *so*, or *pretty*. For example, *very* *good* peaked in 1960 with 1,451 tokens, *very different* in 1870 with 476 tokens, and *very important* in 1960 with 657 tokens. On the other hand, *really good* peaked in 2010, *really sorry* in 2010, and *really great* in 2000. These preliminary findings may point to a replacement of *very* with *really* in an intensifier role over time, a trend noted repeatedly in prior studies. Both *pretty* and *really* appear to be more recent alternatives to *very* in a pre-adjectival position. Unlike COCA, where tokens of *really* outweigh that of *pretty*, *pretty* appears more frequently in COHA. Both intensifiers have peak usage of their most commonly paired adjectives occurring much later than *very*, with *pretty* peaking slightly earlier than *really*. *Pretty*'s earlier peak usage when compared to *really* may help to explain its higher token count, and may also indicate a quite recent replacement with *really* serving as an increasingly common incoming form.

3.2 Results from Reddit analysis

3.2.1 Overall Trends

The total output of the ConvoKit Reddit corpus analysis program resulted in over 21.3 million overall utterances across all of the 29 subreddits that were analyzed. Table 3.5 shows these figures for each category of formality as well as the tokens of *very*, *really*, *pretty*, and *so*.

	Formal	**Neutral**	**Informal A**	**Informal B**	**Total**
Utterances	5,037,974	5,035,549	5,954,388	5,274,796	21,302,707
***very* tokens**	421,214	99,448	244,729	70,423	835,814
***really* tokens**	409,383	196,866	648,203	137,206	1,391,658
***pretty* tokens**	184,407	94,966	266,926	61,514	607,813
***so* tokens**	813,039	338,096	997,477	213,797	2,362,409

TABLE 3.5: Total utterances examined within each category as well as the token count for each intensifier within said categories.

Before analyzing the use of individual intensifiers, a simple calculation was performed to examine if certain categories of subreddits encouraged or discouraged intensification in general. In order to gauge this, the amount of tokens of intensifiers per utterance in each category was calculated. Within this calculation, tokens of *so* were not included, the reasons for which will be discussed shortly. Table 3.6 illustrates the results from this preliminary test.

	Formal	**Neutral**	**Informal A**	**Informal B**	**Total Informal (A+B)**	**Total Overall**
Intensifiers per utterance	0.20147	0.0777	0.19479	0.05102	0.12726	0.1331

TABLE 3.6: Intensifiers per utterance within the different categories of subreddits.

With intensifiers frequently being considered informal aspects of language, it was expected for the more informal categories to have a higher rate of intensification than the formal ones. This, however, was not the case. Instead, the highest rate of intensification was actually found in the Formal subreddits, at 0.20147 intensifiers per utterance, or one intensifier per every 4.96 utterances. This was closely followed by the Informal A category, which had a rate of intensification of 0.19479 intensifiers per utterance, or one intensifier per every 5.13 utterances. Compared to these two categories, the rates of intensification for the Neutral subreddits (0.0777 intensifiers per utterance; one in every 12.87 utterances) and the Informal B subreddits (0.05102 intensifiers per utterance; one in every 19.6 utterances) were quite low. This outcome was somewhat surprising, as the Informal B category was expected to be the most informal, and therefore was expected to see the highest rate of intensification. As will be discussed shortly, the behavior of the Informal B data continues to show inconsistency with predictions.

As a whole, the data showed an overall rate of intensification of 0.1331 intensifiers per utterance, or one intensifier per every 7.51 utterances. While these preliminary findings may be surprising, they may only reveal that the use of intensifiers in general may not have the same connotation of informality within an Internet register as it once held in others. Given that individual intensifiers have their own connotations of formality or informality, closer examination of individual intensifiers' usages within these categories may be more telling of the overall process of intensification on Reddit.

Overall, as well as within each category, *so* was very clearly the most popular of the intensifiers pulled from the subreddits. This is perhaps not too unexpected given the recent upwards trend of its usage, especially in informal registers. However, these figures are also fairly misleading. Given the significant size of the corpus and the amount of data pulled, as well as the limits of the resources used to obtain this data, the program used did not discriminate between instances of *so* functioning as an intensifier and its use as a conjunction, discourse marker, or other speech function. Instead, simply every token of *so* (and each of the other words) was collected. This is not necessarily problematic for *very* and *really* which are used almost exclusively as intensifiers. For *pretty*, tokens of its use as an adjective were almost certainly collected, however a manual glimpse into the utterances from some smaller subreddits indicated that its use as an intensifier outweighs its use as an adjective. Because of this, the following discussion will focus more on *very, really*, and *pretty*. We will return to a discussion regarding the use of *so* later.

After *so, really* was the most common intensifier across the entire sample, outnumbering *very* by 555,844 tokens. This was a predicted outcome given the fact that written internet language is generally more informal and that previous findings show *very* to pattern more so with formal writing. In accordance with this pattern, the formal category of subreddits was the only in which there were more tokens of *very* (421,214) than *really* (409,383). It is important to note that the amount of utterances examined was not exactly equal between each of the four categories. Slightly differing sample sizes means that comparing the amount of tokens of one intensifier in one category to the amount of tokens of that same intensifier in another category is not revealing of its use; rather, differences in its token count may simply be due to more or fewer utterances examined. Instead, analysis focused on the ratio of the amount of tokens of one intensifier to the total amount of intensifier tokens within a specific category or overall.

The percentage of a category's intensifiers constituted by *very* tended to increase with the formality of that category. Making up 41.5% of the intensifiers (specifically tokens of *very*, *really*, and *pretty*) of the Formal subreddits, this percentage dropped to 25.4% in the Neutral subreddits, and 21.1% in the Informal A subreddits. The exception to this trend was in the Informal B category, expected to be the most informal of the four, in which *very* made up 26.2% of all tokens of *very*, *really*, and *pretty*. This trend is reversed for *really*, which makes up 55.9% of the Informal A intensifiers, but only 40.3% of those in the Formal category. Once again, Informal B deviates from the observed pattern. Figure 3.1 shows the percentage of intensifiers made up by *very*, *really*, and *pretty* in each category.

	Formal	**Neutral**	**Informal A**	**Informal B**	**Total**
very	41.5%	25.4%	21.1%	26.2%	27.7%
really	40.3%	50.3%	55.9%	51.0%	50.3%
pretty	18.2%	24.3%	23.0%	22.9%	22.0%

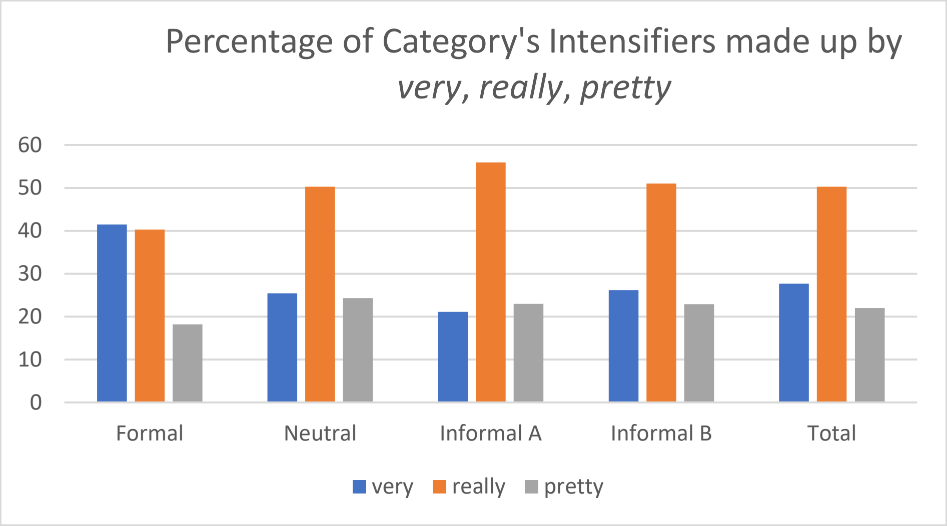


FIGURE 3.1: Within each category, the percentage of the intensifier tokens made up by *very*, *really*, and *pretty* (excluding *so*).

Figure 3.2 specifically highlights how *very* and *really* correspond to more formal and informal comments respectively.

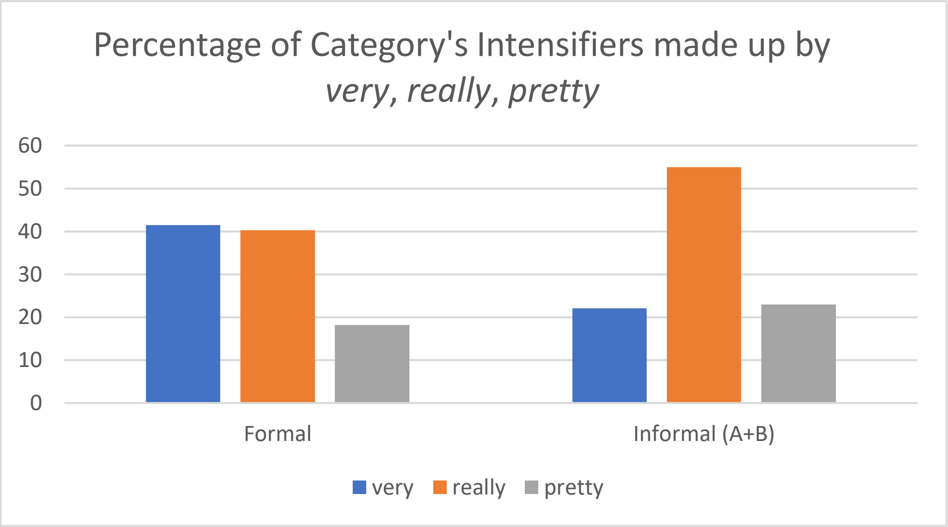


FIGURE 3.2: Percentages of intensifiers made up by *very*, *really*, and *pretty* within the Formal and the two Informal categories.

These figures appear to support the hypothesis that *very* is used more frequently in formal speech and writing, whereas *really* serves as a more informal alternative. That is to say, the use of *very* seems to correspond with the formality of a subreddit, making up a larger percentage of a subreddit's intensifiers when that subreddit is Formal. On the other hand, the opposite appears to be the case for *really*, whose usage corresponds with subreddits that are more informal in nature. With subreddits that are more informal, *really* will constitute a larger percentage of their total intensifiers.

Across the four categories, *pretty* displays a more unique behavior. Its lowest usage is in the Formal category in which it makes up just 18.2% of the intensifier tokens. Its peak usage is found in the Neutral category at 24.3%, and this figure drops as the subreddits become more informal – 23.0% in the Informal A category and 22.9% in Informal B. Despite its more limited presence in the Formal category, *pretty* does not exhibit the same tendency of increased use with increased informality that is seen in the case of *really*. While not necessarily favored in informal writing in the data, *pretty* does seem to be strongly disfavored in subreddits that are more formal in nature.

With an overview of the general patterns noted throughout the data as a whole, what follows is a closer look into each individual category to examine in more detail patterns of intensification and how individual subreddits utilize these adverbs.

3.2.2 Formal Subreddits

The Formal category stands out from the other three due to the prevalence of *very* found within. This category was the only one in which *really* was not the most frequent intensifier. Instead, *very* made up 41.5% of the intensifiers, with *really* trailing slightly at 40.3%. Recall that the criteria for classifying certain subreddits as Formal centered around the topics of discussion and the subreddits' rules surrounding the types of comments allowed. For the most part, this meant that Formal subreddits followed an "AskX" format in which the poster of a topic would present a question and the comments would consist of answers to the question followed by further discussion surrounding that answer. Frequently, rules prohibited commenters from joking or not directly responding to the original poster's (OP) question. r/AskDocs, for example, prohibits top-level comments – comments directly responding to the OP rather than to another comment – from being anything other than an attempt to answer the OP's inquiry. In addition, top-level comments were only permitted for physicians, doctorate level professionals, advanced degree professionals, and medical students, all of whom required verification by the subreddit's moderators.

Rules such as those from r/AskDocs were present in all of the Formal "AskX" subreddits, as well as in r/history, r/linguistics, and r/Physics. Although these three are not constrained to a question-answer format, their rules require on-topic and serious comments rather than jokes or non sequiturs. These rules constraining the type of content to be found in the comments and those who are or are not allowed to comment in the first place may help to explain the significantly higher percentage of tokens of *very* in the Formal category when compared to the others. The conversation topics within the Formal subreddits all center around academic or serious fields with an emphasis on learning, discussing, and responding to questions. Additionally, restrictions on who could participate further narrow the pool of potential commenters to experts in the field or those with relevant knowledge. Given these factors, as well as the explanatory nature of many of the topics present in these subreddits, conversations were likely more serious with a lesser presence of jokes and off-topic statements, and a focus on academic or scholarly writing. The more formal writing, therefore, clearly provided a context for heavier use of *very*.

The strong presence of *very* in the Formal category does not mean that it is absent in more informal categories; over 400,000 tokens were recorded in the Neutral and Informal subreddits. This intensifier certainly maintains a presence in more informal comments. However, the sharp decline in its use outside of the Formal subreddits is noteworthy. Between the non-Formal subreddits, the decline of *very* is much more subtle, making up 25.4% of the Neutral intensifiers and only decreasing to 22.1% of the Informal ones (when considering Informal A and B together). On the other hand, a drop from 41.5% of the Formal intensifiers to 25.4% of the Neutral ones is staggering. While formal situations are not required for *very*, commenters clearly have a connotation with it and formality. This type of connection between the prominence of *very* and a more formal register is not new, and is mirrors a significant finding from Tagliamonte’s (2016) study in which *very* was shown to be the only intensifier to occur in formal writing, with its use declining in more informal emails, texts, and instant messages.

Interestingly, among the Formal subreddits, only two – r/linguistics and r/Physics – had more tokens of *really* than *very*. r/linguistics contained 41,378 tokens of *really* and 40,554 tokens of *very*, and r/Physics had 53,046 tokens of *really* and 48,856 tokens of *very*. The other subreddits, which, with the exception of r/history, were all in the question-answer format, each had more tokens of *very* than *really*. This is perhaps indicative of the preference for *very* in an explanation or lesson, as well as in formal speech overall. This does not detract from the overall strong presence *very* has in the Formal category; clearly the more formal pages are more prone to containing this specific intensifier.

One possible explanation for this pattern may be found in the stances or attitudes held by the commenters on the AskX subreddits towards the conversation. In order to provide an answer to the question posted on these subreddits, one must have the relevant knowledge and experience needed to adequately respond. This creates a type of teacher-student or expert-layman dynamic between the commenter and the OP who asked the question. This difference in power may encourage the “expert” to emphasize their knowledge of the subject with speech that is more formal or academic, therefore including a higher usage of *very* compared to other intensifiers. This perhaps serves as a way of giving legitimacy to their answer and reaffirming their position as “expert” on the matter. Future in-depth study of AskX subreddits may reveal if these types of responses to questions are particularly encouraging to the use of *very*.

3.2.3 Neutral Subreddits

The types of subreddit that are classified as Neutral for this study are the most common type of subreddit found on the website. They can center around any topic about which there is a community willing to converse. This means that Neutral subreddits may revolve around a particular baseball team, hip hop, learning to play a certain instrument, or even sharing videos of cats frantically running around with the “zoomies”. In the case of the present study, topics included the Philadelphia Eagles football team, Spiderman, cooking with cast iron equipment, silly animals, the comic strip *Calvin and Hobbes*, unlikely circumstances, spaces that appear cozy, and particularly photogenic food. Given the immense amount of different topics found on Reddit as well as the more vague qualifications as to what constitutes a Neutral subreddit, this category contained the widest range of conversation topics. This variety is beneficial to maintaining a diverse sample of text, meaning that intensifier use can be examined across a larger variety of contexts.

It is important to note that although these subreddits are considered Neutral when compared to the Formal or Informal subreddits examined, the overall style of Reddit and the internet as a register make the speech found within these pages more informal in nature. In accordance with the patterns that have been revealed thus far, the Neutral subreddits boasted strong favoritism for *really*, accounting for 196,866 tokens, nearly 100,000 more than the runner-up *very*, which produced 99,448 tokens. *Really* made up just over half of all the intensifiers pulled from this category at 50.3%. *Very* and *pretty* had very similar frequencies, making up 25.4% and 24.3% (94,966 tokens) respectively.

With these figures, it is not surprising that *really* was the most used intensifier in each individual Neutral subreddit. When it comes to the second most used, however, the results are more varied. In five of the eight subreddits (r/castiron, r/CozyPlaces, r/AnimalsBeingDerps, r/FoodPorn and r/calvinandhobbes) *very* was used more than *pretty*. In the other three (r/eagles, r/Spiderman, and r/nevertellmetheodds) *pretty* occurred more frequently than *very*. With a similar portion of the Neutral category's intensifiers made up by either one, it is not unexpected to have these mixed results. When taking into account the near equal usage of *very* and *pretty* in this category, as well as the rising use of *pretty* when compared to its low usage in the Formal category, it may be the case that *pretty* is not just emerging as a more informal intensifier, but may be on the way to replacing *very* in informal contexts as *very*'s usage becomes more constrained to formal conversations.

3.2.4 Informal A vs. Informal B

While assigning subreddits to the different categories of formality, it was not initially clear whether or not there would be a significant difference between intensifier usage of the two informal categories. Recall that these categories do not so much differ based on the expected level of formality, but rather on the attitudes held by the subreddits' participants. The Informal A subreddits, despite being casual, encourage legitimate and good-faith discussion and conversation. On the other hand, Informal B subreddits are those in which true discussion of a topic is more infrequent. Instead, jokes, parodies, copypastas[3], and memes that are often crude or inappropriate are the main types of comments found.

An examination of the data from these two categories shows that between the Informal A and Informal B subreddits, the use of *pretty* is nearly identical, constituting 23.0% and 22.9% of the categories respectively. Against expectations, however, the use of *very* noticeably increases (21.1% in Informal A to 26.2% in Informal B) whereas the use of *really* decreases (55.9% in Informal A to 51.0% in Informal B). Given the stagnation of *pretty* in this context, the decline in *really*'s use appears to be compensated by *very*'s increased occurrences. Despite this pattern, *really* still clearly outweighs *very* in both categories, reaffirming its position as the favored intensifier. Even with the change in the rates of usage of *very* and *really* between Informal A and B, when accounting for both Informal categories together, the overall trend of decreasing use of *very* with decreasing formality is still visible, as is the opposite trend for *really*. Figure 3.3 shows the percentage of intensifier tokens made up of *very, really*, and *pretty* in both Informal categories together compared to the Neutral and Formal categories.

	Formal	**Neutral**	**Informal A+B**	**Total**
very	41.0%	25.9%	22.1%	27.7%
really	40.2%	50.0%	55.0%	50.3%
pretty	18.8%	24.1%	23.0%	22.0%

[3] A copypasta is a type of written meme that circulates generally due to its funny or absurd nature. Ranging from a few sentences to multiple paragraphs, certain specific aspects of the copypasta may be substituted for something more directly related to the overall conversation, while the rest of the meme remains the same as it spreads.

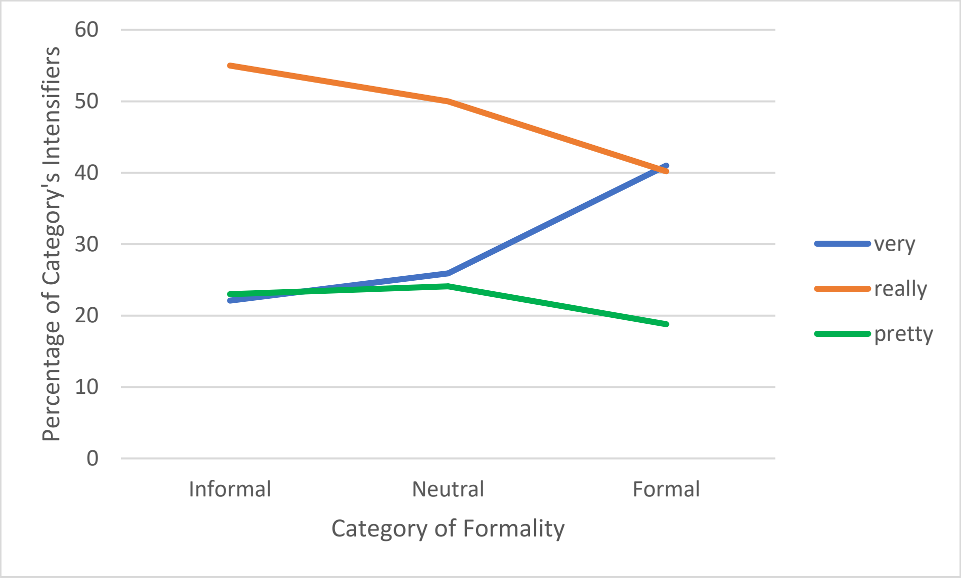


FIGURE 3.3: Percentages of intensifiers made up by *very*, *really*, and *pretty* within the Informal, Neutral, and Formal categories.

The results from the Informal B category were surprising in a number of ways. For one, given the assumption that the subreddits in this category were the most informal, the percentage of the category’s intensifiers made up of *very* and *really* were expected to be the lowest and highest respectively when compared to the other categories. Instead, it exhibited the second highest percentage of *very* (behind Formal) and the second highest percentage of *really* (behind Informal A).

Another reason this category was intriguing was due to its relatively low rate of intensification overall, especially when compared to that of the other categories. In the Informal B subreddits, there was one intensifier about every 19.6 utterances, by far the lowest rate of the four categories. When looking at the data from all categories of subreddits together, the results from Informal B act like an outlier within an otherwise clear trend. The use of *very*, for example, shows a clear decline as the formality decreases, peaking in usage in the Formal category, showing a lower rate in Neutral, and reaching its lowest usage in Informal A. However, in Informal B, the rate of *very* rebounds, exceeding that of the Neutral subreddits. Likewise, the overall data shows an increase in usage of *really* as formality decreases, hitting its lowest rate in Formal, increasing in Neutral, and peaking in Informal A. Once again, Informal B data disrupts this trend, exhibiting a lower rate of usage of *really* than that found in Informal A.

Why, then, does the Informal B data digress from these patterns? One possible explanation may be the type of comments made within these subreddits. Much of the content on these subreddits may be considered "shitposting", which describes a type of engagement where participants are purposefully foolish, derogatory, obtuse, and satirical. Good-faith discussion is not usually expected in these contexts, including in the subreddit r/shitposting which is specifically centered around this type of content. In other subreddits, specifically of the "circlejerk" variety, conversations are frequently satirical in an effort to parody conversations seen elsewhere. The subreddit r/GamingCircleJerk, for example, provides a space for its users to mock posts, images, and opinions found in other gaming-related subreddits.

Much of the conversation from the Informal B subreddits are almost performative, albeit frequently in a crude manner. These subreddits, therefore, may not be an ideal source of natural, everyday conversation in the same way that those in the Informal A category are. This may help to explain why the results from Informal B stand out. This is to say, while the behavior within the Informal B subreddits is certainly not formal in nature, it may also not be an accurate depiction of actual informal conversation, but may rather be more representative of intentionally obtuse, performative satire and parody. However, before this conclusion is accepted as true, further study that specifically examines the language in "shitposting" subreddits is certainly warranted. Additionally, future inclusion of *so* in an analysis of these subreddits may also change our understanding of how intensifiers are used.

3.3 More about *so*, and further intensifier testing

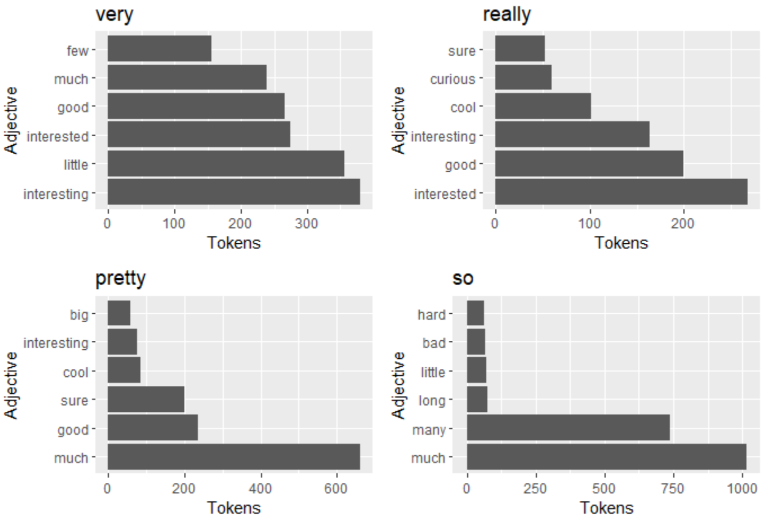
3.3.1 Preparing a second round of testing

Due to the limitations preventing detailed analysis on the use of *so* in the data taken from Reddit, the large amount of tokens of this particular intensifier that were recorded were not necessarily useful. Therefore, a second round of testing was introduced in order to attempt to better understand how *so* (as well as the other intensifiers) were being used in the data, specifically before adjectives. In order to achieve this, another Python script was written that would pull 25,000 utterances from the largest subreddit in each of the four categories – r/history, r/eagles, r/CasualConversation, and r/circlejerk. Importantly, the program only collected utterances that contained a token of *very*, *really*, *pretty*, or *so*, forming a total of 100,000 utterances. Within this pool of 100,000 utterances, a second Python script used the Natural Language Toolkit (NLTK) package to tag each word with its part of speech. Any time the program would see an instance of an intensifier directly followed by an adjective (tagged 'JJ'), the intensifier-adjective bigram would be appended to a spreadsheet to visualize trends of the frequencies in which these intensifiers emphasized adjectives, as well as the specific adjectives that paired with each intensifier.

Because of a significantly more constrained sample size, hardware limitations (while still present), were less of a hinderance, allowing for this more detailed analysis. A potential issue with this test comes with how the scripts identify certain words and their parts of speech within the data. For example, one intensifier-adjective bigram that resulted was *very subject*. Certainly, *subject* can function as an adjective in a sentence such as "The northeast region of the country is very subject to blizzards." However, it can also be a noun, and in the data, it indeed serves this function: "…this very subject has kept me awake many a night." Small errors like these result from the functionality of NLTK's part of speech tagger, and avoiding these instances is extremely difficult. The vast majority of the data collected and analyzed in this second round of testing is fine, however it is important to keep these details in mind when discussing the results.

For the purpose of the current study, the discussion of this round of analysis will focus on data from the Formal subreddit (r/history) and the Informal A subreddit (r/CasualConversation). As explained earlier, the language in the Informal B category is highly performative and not necessarily indicative of actual casual conversation, as is illustrated in its strong deviations from noted trends in the rest of the data. For this reason, the largest Informal A subreddit was chosen to represent informal intensifier use of Reddit for this test, and the data from the Formal subreddit was used to allow for comparisons between the two extremes of the established formality spectrum. Figures 3.4 and 3.5 show the six most frequent intensifier-adjective bigrams for each of the four intensifiers (*very*, *really*, *pretty*, and now *so*) for the Formal subreddit and Informal A subreddit.

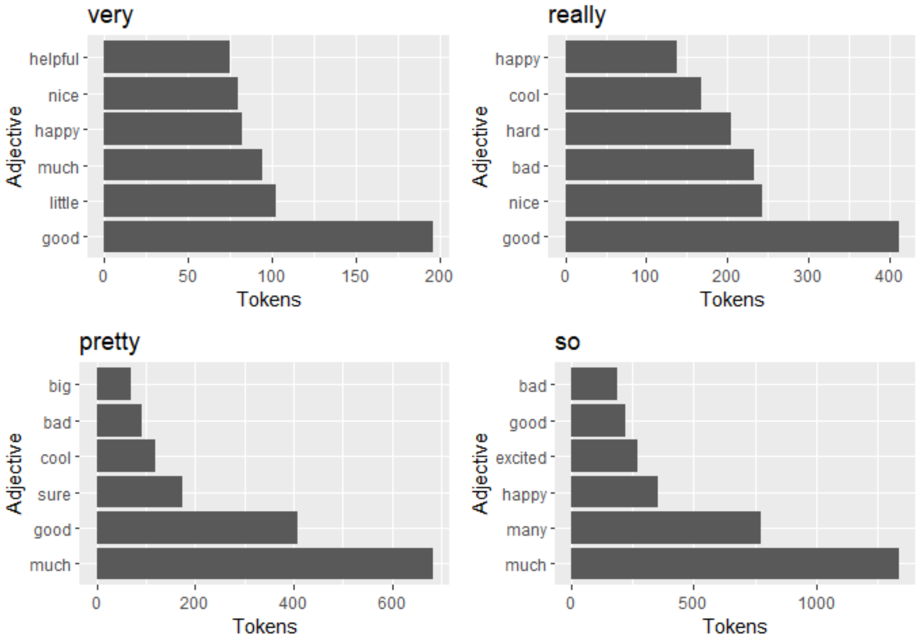
Formal intensifier-adjective Bigrams



very		**really**	
Bigram	**Tokens**	**Bigram**	**Tokens**
very interesting	380	*really interested*	268
very little	357	*really good*	200
very interested	275	*really interesting*	164
very good	266	*really cool*	101
very much	239	*really curious*	60
very few	155	*really sure*	52
pretty		**so**	
Bigram	**Tokens**	**Bigram**	**Tokens**
pretty much	662	*so much*	1,015
pretty good	237	*so many*	738
pretty sure	201	*so long*	75
pretty cool	85	*so little*	70
pretty interesting	76	*so bad*	65
pretty big	59	*so hard*	63

FIGURE 3.4: Most frequent intensifier-adjective bigrams from the Formal sample.

Informal intensifier-adjective Bigrams



very		**really**	
Bigram	**Tokens**	**Bigram**	**Tokens**
very good	196	*really good*	411
very little	102	*really nice*	242
very much	94	*really bad*	232
very happy	82	*really hard*	204
very nice	80	*really cool*	168
very helpful	75	*really happy*	138
pretty		**so**	
Bigram	**Tokens**	**Bigram**	**Tokens**
pretty much	683	*so much*	1,341
pretty good	407	*so many*	777
pretty sure	175	*so happy*	354
pretty cool	118	*so excited*	272
pretty bad	92	*so good*	219
pretty big	70	*so bad*	187

FIGURE 3.5: Most frequent intensifier-adjective bigrams from the Informal sample.

3.3.2 Results for formal bigrams

Of the 25,000 Formal utterances containing a token of *very*, *really*, *pretty*, or *so*, 19,290 intensifier-adjective bigrams were identified. This means that there were 5,710 tokens of these adverbs in which they did not function as intensifiers according to the NLTK part of speech tagger. Of these 19,290 bigrams, 8,031 involved the use of *so*, seemingly making it the most common of the four intensifiers within this sample. However, the NLTK part of speech tagger erroneously tagged the pronoun *I* as an adjective 2,906 times; therefore, instances of *so I* bigrams were omitted. This meant that there were actually 5,125 instances of *so* as an intensifier from the 25,000 utterances from r/history. Table 3.7 shows the token counts for each of the four intensifiers within this 25,000-utterance sample.

Intensifier	**Tokens**
very	6,230
so	5,125
pretty	2,755
really	2,274

TABLE 3.7: Token counts for each intensifier within the 25,000-utterance sample from r/history.

With the ability to now examine how *so* functions as an intensifier within Reddit, we can see that it is quite frequent, even in a formal subreddit. Given the general informality of the Internet as a register, higher use of *so* was expected, and this certainly appears to be the case within the r/history sample. Still, *very* maintains its position as the favored formal intensifier, and *really* trails the group, falling behind even *pretty* as the least used intensifier. This once again indicates a preference for *very* over *really* in more formal conversations on Reddit.

An examination of the adjectives used with each intensifier reveals how each may be utilized by commenters. For instance, *so* has a substantial amount of tokens within this sample, however its two most frequently-paired adjectives – *much* (1,015 tokens) and *many* (738 tokens) – significantly outweigh the third adjective, *long*, which accounts for just 75 tokens. The bigram *so long*, however, can be a bit problematic as it can function as one unit to serve as an expression of farewell. Therefore, it is possible that not all of the instances of *so long* involve the emphasis of the length of some noun. The adjective *little* (70 tokens), then, becomes the next most frequent where its use with *so* can safely assume to be intensified. The discrepancy between the instances of *so much* and *so many* and the use of *so* with other adjectives is staggering. These results indicate that *so* often functions as an intensifier to emphasis quantity, rather than just the degree to which an adjective describes something. Approximately 34.2% of the tokens of *so* in the r/history sample were used in this way.

When comparing the *so* bigrams with the results from the COCA test, we can see that *bad* and *hard* appear among the most frequently used adjectives in both samples. However, this is not too revealing about *so*, as these are common, simple, adjectives that occur frequently with other intensifiers within the COCA data as well. Interestingly, the most common adjectives that occur with *really* do not match those found from the COCA test, with the exception of *good*. Instead, we have some more complex adjectives, including *interesting*, *interested*, and *curious.* There does not seem to be as obvious of a discrepancy between the adjectives used with *really* and those used with *very* within this r/history sample, even though more complex ones may be expected to pair with the latter. Based on the similarity between the adjectives used with either intensifier as well as the significant difference between the quantities of each intensifier, when it comes to expressing formality within this Formal subreddit, doing so may be less reliant on the adjective used, but rather more so on the choice of intensifier.

Notably, the distribution of adjectives used with *so* and *pretty* shows just one or two favored choices with these intensifiers. On the other hand, *very* and *really* have a much more equally dispersed distributions of the adjectives that pair with them. The most commonly used adjective with *very* was *interesting* (380 tokens, and the seventh most frequent in COCA), but the second most commonly used adjective *little*[4] (357 tokens) is not far behind. The 275 tokens of *interested* and 266 tokens of *good* which follow help to illustrate this more even distribution of adjectives that can be used with *very*. *Interested* and *interesting* also make up two of the most common adjectives used with *really*, at 268 tokens and 164 tokens respectively – and show this type of even distribution as well. Falling between these two is *good* with 200 tokens, and *cool* is the fourth placed adjective with 101 tokens.

The differences in adjective distribution for the four intensifiers may reveal how each is used as well as how each has been (or is being) grammaticalized. Based on this bigram analysis, *very* and *really* primarily occur in an intensifier-adjective position specifically in an intensifying function. Despite paring widely with many different adjectives, the relationship between these two intensifiers and their adjectives appears to be purely emphatic. In contrast, *pretty* and *so* exhibit more than one type of use when occurring in intensifier-adjective bigrams.

The use of *so* appears to frequently intensify quantity in addition to degree, and *pretty* patterns strongly with *much, good,* and *sure*, forming frequent, non-compositional expressions that can be found in everyday speech. Both *pretty* and *so* can also function as typical intensifiers, emphasizing degree adjectives, however they are not restricted to this role like *very* and *really*

[4] The *very little* bigrams may be misleading, as *little* can be used as a noun in this context, as in the sentence “There is very little left of ruins.” It is possible that some instances of *little* functioning as a noun may have been incorrectly tagged as an adjective by NLTK.

appear to be. All four degree adverbs do serve as intensifiers, but the data indicates that they do not share the same functional distribution.

Recall that Ito and Tagliamonte (2003) used the distribution of *very* and *really* to posit differing degrees of grammaticalization for each, determining that *very* was further along than *really* due to it more heavily favoring a predicative over attributive position. While both adverbs shared this preference, *very*'s preference for a predicative position was much stronger than that of *really*, pointing to its further progress in the process of grammaticalization. A similar pattern is perhaps becoming visible in our current data, showing a more restricted function for *really* and *very* (emphasis) when compared to *pretty* and *so* (emphasis, quantity, and non-compositional expressions). Ito and Tagliamonte explain *very*'s more advanced position further, citing Partington (1993:183), who explains that a greater "width of collocation" (i.e., having a greater distribution of adjectives to pair with) may also reveal further grammaticalization. Indeed, our data also shows that *very* and *really* have a larger and more equal distribution of the adjectives found to pair with them, perhaps due to their more specialized function. While these patterns do seem to indicate *very* and *really* being further grammaticalized than *pretty* and *so*, our data is not quite sufficient to claim this further progress. Instead, it appears to better reveal that both pairs of intensifiers are perhaps following similar, but separate, paths of grammaticalization in which intensification is present in both, but in which *pretty* and *so* adopt additional functions.

Whereas *very* and *really* have reached a point where their primary function is the emphasis of adjectives, *pretty* seems to have adopted a different function that has led to the formation of a handful of common expressions that, although look like intensifier-adjective bigrams, do not necessarily function as such. *Pretty much*, for example, does not emphasize quantity in the same way that *so much* does; instead, it is a fixed expression of affirmation. On the other hand, the primary intensifier function of *so* tends to emphasize quantity over quality. Its intensification of the latter is still present in the data, but the distribution of its use in the sample heavily favors its appearances with *much* and *many*. Indeed, both *pretty* and *so* exhibit behavior of grammaticalized intensifier forms like *very* and *really*. However, the lopsided distribution of the adjectives that are found to occur with them indicates unique paths of grammaticalization. While preliminary, these findings may give way to further research regarding the grammaticalization of *pretty* and *so*.

3.3.3 Results for informal bigrams

From the 25,000 utterances pulled from the Informal subreddit r/CasualConversation, 29,166 intensifier-adjective bigrams emerged. This indicates that there were over 4,000 utterances in which more than one of these bigrams appeared. However, just like with the Formal bigrams, all instances of *so I* had to be removed due to NLTK's incorrect part of speech tagging. There were 7,942 *so I* bigrams, leaving behind 21,224 total intensifier-adjective bigrams from the Informal category, nearly 5,000 more than those in the Formal category. This larger quantity may hint at a stronger tendency for informal commenters to use intensifiers in general, however we still see a slightly higher intensifier per utterance ratio in the overall data from the Formal category (see Table 3.6). Table 3.8 shows the most frequent intensifiers from this Informal sample.

Intensifier	**Tokens**
so	8,608
really	5,091
very	3,958
pretty	3,567

TABLE 3.8: Token counts for each intensifier with the 25,000-utterance sample from r/CasualConversation.

Whereas *very* was the favored intensifier from the Formal sample, it has taken a lower position in the Informal sample, trailing *really* and *so*, the most frequent. The increase in the amount of tokens of *very* from the Informal category (3,958 tokens) to the Formal category (6,230 tokens) shows a growth of about 57.4% and a correlation between its use and more formal conversation. On the other hand, the amount of tokens of *really* increases from 2,274 tokens in the Formal category to 5,091 tokens in the Informal category, growing 123.9%. These figures corroborate the patterns discussed earlier in which the frequencies of tokens of *very* and tokens of *really* increase with more formal and more informal conversations respectively. These patterns support prior findings and opinions that link *very* to formal speech (Stoffel 1901, Fries 1940, Tagliamonte 2016). The decrease in tokens of *very* from formal to informal conversation seems to be compensated by moderate increases in token counts for *really* and *pretty*, as well as a substantial increase in token counts for *so*.

With the uses of *so* limited to just the intensifier function in this second round of testing, we are now able to see how its usage changes with formality. It certainly appears more frequently in the Informal category (8,608 tokens) than in the Formal one (5,125 tokens), showing an increase of about 68%. Despite a fairly significant difference in amounts of tokens of *so*, a glance at the distribution of adjectives it emphasizes shows that its function in the Informal category is quite similar to that in the Formal one.

As with the Formal category, the Informal uses of *so* show that the *much* (1,341 tokens) and *many* (777 tokens) are easily the most frequent. The next most frequent adjectives used with *so* are *happy* (354 tokens) and *excited* (272 tokens). In both categories, therefore, *so much* and *so* *many* are clear favorites when it comes to intensification. Although these two bigrams heavily outweigh the others, the distribution of adjectives in the Informal sample is much less lopsided than in the Formal one. We can see that subsequent adjectives such as *happy* and *excited* still appear rather commonly, and with much greater frequencies than *long* and *little* in the Formal sample. With a greater frequency of *so* bigrams and a more even distribution of adjectives in the Informal sample, it appears that *so*'s usage as an intensifier is more prominent in informal conversations.

Interesting, *pretty* behaves similarly in both categories, albeit with about 800 more tokens in the Informal sample. In both categories, the four most common adjectives used with *pretty* are (in descending order of frequency) *much, good, sure,* and *cool*, with the latter three being the most common adjectives from the COCA test. Table 3.9 shows the frequencies of these four bigrams in both the Formal and Informal samples.

FORMAL		**INFORMAL**	
Bigram	**Tokens**	**Bigram**	**Tokens**
pretty much	662	*pretty much*	683
pretty good	237	*pretty good*	407
pretty sure	201	*pretty sure*	175
pretty cool	85	*pretty cool*	118
Total	**1,185**	**Total**	**1,383**

TABLE 3.9: Distribution of most common *pretty* bigrams in Formal and Informal samples.

With the exception of *pretty sure*, each bigram increases in frequency from the Formal to the Informal sample, with *pretty good* exhibiting the largest increase (170 tokens). While it appears that *pretty* is favored in more informal conversations, the way it is used as an intensifier does not seem to change based on formality. Instead, the most popular ways to use *pretty* seem to involve the employment of a select few common expressions.

Once again, *very* and *really* show a more restricted context of use than *so* and *pretty*, occurring primarily as degree adverbs and illustrating how both pairs may be on different paths of grammaticalization into intensifiers. Possibly due to this narrower context, the distribution of adjectives paired with *very* and *really* is more evenly dispersed than that of the adjectives paired with *pretty* and *so.* The adjective *good* is the most frequent in the Informal sample, accounting for 196 tokens with *very* and 411 with *really*. Still, the next five most frequent adjectives do not trail too far, and their distributions relative to each other are quite similar, as is shown in Table 3.10.

***very* Bigrams**		***really* Bigrams**	
Bigram	**Tokens**	**Bigram**	**Tokens**
very little	102	*really nice*	242
very much	94	*really bad*	232
very happy	82	*really hard*	204
very nice	80	*really cool*	168
very helpful	75	*really happy*	138
Total	**433**	**Total**	**984**

TABLE 3.10: The most frequent *very*-adjective and *really*-adjective bigrams from the Informal sample.

3.3.4 Summarizing the Findings: Intensifiers, formality, and grammaticalization

Overall, we can see that *very* and *really* intensify a wider range of adjectives in more evenly dispersed frequencies, regardless of formality. On the other hand, *pretty* and *so*, while still able to act as intensifiers for various common adjectives, exhibit a strong tendency to more frequently pair with a select few adjectives, thereby producing a smaller subset of *pretty*-adjective and *so*-adjective bigrams that appear much more frequently than the rest. This discrepancy between how *very* and *really* appear and how *pretty* and *so* appear possibly illustrates the former two being further along in the process of grammaticalization into intensifiers, and having a context of application that is more constrained to the emphasis of degree of an adjective. Given the recency associated with *so*'s increased use as well as the fact that the non-grammaticalized form of *pretty* (an adjective meaning "attractive, physically appealing") still occurs in English, indications that they are not as far along the path of grammaticalization are not unexpected. Furthermore, *pretty* and *so* find themselves to seemingly be undergoing separate tracts of grammaticalization, forming common discourse-marker expressions in the case of the former and expressions of emphatic quantity in the latter.

In terms of formality, the frequencies in which the four intensifiers occur in the Formal and Informal samples also show how the choice of intensifier may pattern with different contexts. Again, *very* is the most commonly used intensifier in the Formal sample, while dropping behind *so* and *really* in the Informal sample. *So, really*, and *pretty*, on the other hand, reveal noticeable increases in their usage in the Informal category, with *so* being a particularly frequent intensifier. Therefore, the data from this analysis as well as from the larger Reddit corpus analysis supports previously noted patterns that *really* and *so* are more informal intensifiers when compared to *very*, which is more indicative of formal speech.

4. Discussion

4.1 Intensifiers and Formality

The association of intensifiers or degree adverbs with informal speech is not a new phenomenon, rather it has been remarked upon for over a century. *Very*, however, has consistently been considered an outlier – marking formal speech or writing. The recent emergence of the Internet as a new register for relatively informal written language provides an important opportunity to gauge how these forms change or maintain their previously noted indexicality. Given the expected overall informal setting of Reddit, the prominence of *really* as the most frequent of the examined intensifiers supports the reported findings of it being the popular informal or “vulgar” choice. Likewise, the maintenance of *very* as the most frequent intensifier in the Formal subreddits, as well as its decline in usage with more informal conversations, show that its categorization as a formal or “standard” intensifier is accurate and that this classification transcends the medium of conversation.

4.2 Limitations

While a significant amount of data was collected and analyzed in this current study, there were certainly limitations and room for improvement. One such possible issue is the subjectivity associated with the selection of subreddits that were studied. The categorization of these pages into the four categories of formality were of my own discretion based on a number of factors contributing to how these subreddits operate. Someone recreating this study or a similar one may decide to categorize the subreddits differently, which may lead to varied results. One may decide that splitting the Informal subreddits into an A and B category is not necessary, whereas another researcher may decide to disregard Informal B subreddits altogether.

In the current study, several decisions were made in order to limit the scope of the data that was collected, generally because of hardware limitations. The ConvoKit Reddit corpus includes 948,169 subreddits, yet only 29 were used in this investigation. While the roughly 21 million utterances scraped in this project is no small amount, other corpus-based studies can more efficiently gather much larger quantities, creating a fuller sample that is more representative of the corpus as a whole. Ideally, a future iteration of this study would allow for collection of data from the entire Reddit corpus in order to paint a more detailed picture of the language used on the website as a whole.

Another notable aspect that could be improved is the way limitations of the hardware used for the Python scripts as well as my own knowledge of computational linguistic analysis impacted the data scraping process. Although I was able to locate and utilize resources to aid in my learning of Python and R, the scripts could have likely been more efficient in terms of data collection and organization. Cleaner scripts could lead to the ability to restrict the data in each category to exactly the same amount of utterances and subreddits. Instead, the utterance counts in each category were close, but not exactly equal, and these categories differed in the amount of subreddits needed to reach the five million utterance goal. A future study with equal utterance and subreddit counts will have a cleaner sample size and an equal distribution of topic diversity. Additionally, the inability to include *so* in the first Reddit analysis may mean that the collected data does not completely represent the full picture of intensifier use. The relative frequencies of *very*, *really*, and *pretty* are revealing of some preferences among speakers, however it is possible that the inclusion of *so* in this analysis would result in slightly different patterns. Ideally, a continuation or recreation of this study would include all four intensifiers in the main analysis.

Finally, the scripts used had some limitations based on how they searched for certain words in the corpus and tagged their parts of speech, such as the example in which *subject* was tagged as an adjective despite functioning as a noun. The ability to clean up instances like this in future tests would be very difficult, but extremely valuable nonetheless.

4.3 Directions for future study

The findings from the current study are certainly indicative of how intensifiers may be used in an informal written register, as well as how the overall usage of these intensifiers may be changing over time. Of course, there is much yet to be discovered about these forms. Future study warrants further consideration as to which subreddits fit in which formality category, as well as if there is a more adequate way to categorize based on formality in the first place. Interesting patterns of intensifier usage can be found within individual categories, notably the Formal and Informal B ones. With discrepancies between the question-answer and general discussion Formal subreddits, as well as the satirical and performative language found in the Informal B subreddits, future research may find value in taking the commenters' stances towards the topic into account more. Although discussed briefly above, speakers' stances and attitudes may be worth exploring further to help explain some of the more unique patterns identified in these two categories.

In addition to stance, taking into account commenters' socioeconomic characteristics may help to reveal patterns of intensifier use with gender, age, and education. While this may be challenging due to Reddit users' anonymity, subreddits designed for participation of specific genders or ages exist and may allow for the examination of these types of patterns. The current investigation can serve as a solid baseline analysis of how intensifiers are used on Reddit; however, the website's demographics suggest that the sample utilized here is predominantly made up of young males. Further use of Reddit as a source of data would benefit from the utilization of subreddits intended for specific audiences, allowing for dimensions of gender, race, age, and socioeconomic status to be considered.

While the pre-adjectival position for intensifiers is quite common, it does not capture every environment where an intensifier may occur. Stacked usage (e.g., *really really good*), for instance, may also be worth studying in order to fully understand variation among intensifier use. An extension of this type of study warrants inclusion of stacked intensifiers, which may even help to explain why certain pairs of stacked intensifiers (such as *really really* and *so very*) are acceptable whereas others (**very really*, **so pretty*, etc.) are not.

The function of *pretty* as an intensifier is also quite interesting and may be worth studying in more detail in the future. Its uniqueness comes from its ability to emphasize an adjective like any typical intensifier, but to also qualify or temper an adjective. Differences in its usage seem to be context dependent as well as distinguishable based on a speaker's inflection. A falling intonation from *pretty* to the following adjective may indicate a qualifying role, whereas a rising or unchanging intonation may show the emphatic intensifier function. For example, a common response to the question "How are you doing?" may be "Pretty ↘ good," indicating that the responder is good, but not too good. On the other hand, a statement like "This chocolate cake is pretty ↗ good!" emphasizes the quality of the cake, implying that it is better than just good. Of course, these observations are partially anecdotal, but further investigation of these patterns may help to more thoroughly describe the unique behavior of the adverb *pretty*.

Additionally, there may be value in revisiting this project with more powerful hardware and scripting capabilities. A more efficient Python script would ideally be able to only take instances of the adverbs when they occur in a pre-adjectival intensifier role. While *really*, for example, can arguably serve as an intensifier for a verb (e.g., *I really enjoy the smell of coffee*), the present study focused on intensifier-adjective structures. The first round of analysis was unable to specifically parse out these bigrams, and therefore a second, more restricted test was needed. Ideally, any pre-adjectival intensifier would be able to be analyzed from the original 21.3 million utterances that were collected, allowing for a more straightforward and insightful examination on how the use of one specific adverb compares to another. In such a case, instances of *really* intensifying a verb or *pretty* functioning as an adjective would be avoided, and comparing the frequency of each intensifier would be significantly easier, especially in the case of *so*.

Finally, in studying intensifier usage overall, not necessarily confined to use on the Internet, there may be value in considering Eckert's (2008) concept of the indexical field. Connotations between certain intensifiers and formality or informality have been noted repeatedly in the past, and have been supported in the present study. The results of this study, for example, illustrate an association of *very* with more formal writing, and a specifically strong use in educated, explanatory, and scholarly conversation. By considering the indexical field, it may be possible to note how the formal association of *very* may also have ties with education, and therefore expertise and even power over one with less knowledge. Likewise, given previous studies' findings associating a higher use of *very* with age, this potential field of indexicality for *very* may expand to include older age and maturity.

While the Internet is no longer in its infancy, it is still a very new medium of communication that serves, and will continue to serve, as an enormous collection of language. Therefore, we should be cognizant of how patterns noted in oral or written language may surface or change when they enter cyberspace. Many aspects of any given language are always changing, and intensifiers do so relatively rapidly. Examining their usage on the Internet may therefore reveal how this register may encourage or discourage linguistic variation – the results could be very intense.

5. References

Biber, D., & Finegan, E. (1988). Adverbial stance types in English. *Discourse Processes*, *11*(1), 1–34. https://doi.org/10.1080/01638538809544689

Dean, B. (2023, March 27). *Reddit user and growth stats (updated March 2023)*. Backlinko. https://backlinko.com/reddit-users

Eckert, P. (2008). Variation and the indexical field 1. *Journal of sociolinguistics*, *12*(4), 453-476.

Fries, C. C. (1940). *American English grammar: The grammatical structure of present-day American English with especial reference to social differences or class dialects* (No. 10). D. Appleton-Century Company, incorporated.

Hopper, P. J. (1991). On some principles of grammaticization. *Approaches to grammaticalization*, *1*, 17-35.

Hopper, P. J., & Traugott, E. C. (2003). *Grammaticalization* (2nd ed). Cambridge University Press.

Ito, R., & Tagliamonte, S. (2003). Well weird, right dodgy, very strange, really cool: Layering and recycling in English intensifiers. *Language in Society*, *32*(2), 257–279. https://doi.org/10.1017/S0047404503322055

Jespersen, O. (1922). *Language: Its nature, development and origin*. London : G. Allen & Unwin ltd. http://archive.org/details/afa5370.0001.001.umich.edu

Kim Jong-Bok, & Grace Gesoon Moon. (2014). The SKT construction in English: A corpus-based perspective. *Linguistic Research*, *31*(3), 519–539. https://doi.org/10.17250/KHISLI.31.3.201412.005

Levin, H., Long, S., & Schaffer, C. A. (1981). The formality of the Latinate lexicon in English. *Language and Speech*, *24*(2), 161-171.

Lewis, D. (2020). Speaker stance and evaluative -ly adverbs in the Modern English period. *Language Sciences*, *82*, 101332. https://doi.org/10.1016/j.langsci.2020.101332

Lühr, R. (1984). Reste der athematischen Konjugation in den germanischen Sprachen. *Das Germanische und die Rekonstruktion der indogermanischen Grundsprache*, 25-90.

Meillet A. (1912). *L' évolution des formes grammaticales*. N.Zanichelli ; F.Alcan ; Williams et Norgate.

Méndez-Naya, B. (2003). On Intensifiers and Grammaticalization: The Case of SWIþE. *English Studies*, *84*(4), 372–391. https://doi.org/10.1076/enst.84.4.372.17388

Molina, B. (2017, August 31). *Reddit is extremely popular. Here’s how to watch what your kids are doing*. https://www.usatoday.com/story/tech/talkingtech/2017/08/31/reddit-extremely-popular-heres-how-watch-what-your-kids-doing/607996001/

Mustanoja, T. F. (1960). *A Middle English syntax: Parts of speech* (Vol. 23). Société néophilologique.

Partington, A. (1993). Corpus evidence of language change. *Text and technology. In honour of John Sinclair*, 177-192.

Peters, H. (1994). Degree Adverbs in Early Modern English. In *Studies in Early Modern English* (pp. 269–288).

Quirk, Randolph; Greenbaum, Sidney; Leech, Geoffrey; & Svartvik, Jan (1985). A comprehensive grammar of the English language. New York: Longman

Stoffel, Cornelis (1901). Intensives and down-toners. Heidelberg: Carl Winter.

Tagliamonte, Sali. 2004. “Intensifiers in Toronto.” Unpublished MS.

Tagliamonte, S., & Roberts, C. (2005). SO WEIRD; SO COOL; SO INNOVATIVE: THE USE OF INTENSIFIERS IN THE TELEVISION SERIES FRIENDS. *American Speech*, *80*(3), 280–300. https://doi.org/10.1215/00031283-80-3-280

Tagliamonte, S. (2016). So sick or so cool? The language of youth on the internet. *Language in Society*, *45*(1), 1-32. doi:10.1017/S0047404515000780

6. Vita

Joshua Baumgarten was born on May 7, 1997 in Abington, Pennsylvania. He was raised in Yardley, Pennsylvania, before attending college at the University of Pittsburgh. He graduated in 2019 with a B.A. in Linguistics and Spanish, as well having minored in Portuguese and Luso-Brazilian culture.

He then attended Syracuse University where he received an M.A. in Linguistics with a concentration in Language, Culture, and Society in June of 2023. At Syracuse Universtiy, he was the instructor of record for multiple Spanish courses, obtained a Certificate of University Teaching, and received an award for Outstanding M.A. Student in Linguistics.