AN ABSTRACT OF THE DISSERTATION OF

Christopher McBride for the degree of Doctor of Philosophy in Counseling presented on February 28, 2022.

Title: Linguistic Characteristics of Health Anxiety from Online Discourse

Abstract approved: _____
Kok-Mun Ng

Health anxiety prevalence is increasing (Tyrer et al., 2019). Health anxiety is a preoccupation or overestimation of the serious physiological symptoms that cause significant distress (Salkovskis & Warwick, 2001). Those with health anxiety may endure significant suffering, intense misery, and for some health anxiety leads to suicide (Tyrer & Tyrer 2018). It is a condition that is often under diagnosed and the average patient may go years prior to appropriate diagnosis (Hedman et al, 2011; Tyrer & Tyrer 2018). The study of language has long history within the field of mental health. Language use can provide information about an individual's beliefs, social relationships, personality, thinking patterns, and fears (Pennebaker et al., 2015). Additionally, language can be investigated to identify important psychological markers that are indicative of an individual's inner workings (Choundry et al., 2013). Little is known about the linguistic attributes specific to health anxiety in online discourse. To address the research gap, two studies were conducted to determine the linguistic attributes of individuals with health anxiety. The first study used the Language and Inquiry Word (LIWC) application to evaluate categorical data for online health anxiety communication (Pennebaker et al., 2015). The second study utilized AntConc to identify keywords and collocations associated to identify what words make health anxiety discourse distinct from other forms of online communication (Anthony, 2020).

The first study examined the summary, linguistic, and psycholinguistic frequencies of word usage for health anxiety communication. One year of posts and comments were extracted from the subreddit r/HealthAnxiety to create a study corpus. The Corpus of Contemporary American English was used as a reference corpus (COCA; The COCA, 2021). The research questions were:

1. What is the score of summary variables about health anxiety?
2. What is the level of use of linguistic processes in online posts about health anxiety?
3. What is the pattern of use of linguistic processes variables in online posts about health anxiety compared to a reference corpus?
4. What is the level of use of psychological processes in online posts about health anxiety?
5. What is the pattern of use of psychological processes in online posts about health anxiety compared to a reference corpus?

Descriptive statistics were reported related to scores and level of use in the study corpus. For the study corpus, summary variables scores indicate that those in the study corpus are high in authenticity, low in emotional tone, low in analytic processes, and low in clout. For linguistic and psychological process variables, log-likelihood ratio (G^2) and Bayes Information Criterion (*BIC*) were used to compare the study and reference corpus. Log-likelihood for all variables understudy exceeded the critical value for significance (G^2 = 484579.4 to 276.0733, df = 1, $p <$ .01). Bayes factor (*BIC*) scores for results were "very strong" (*BIC* = 465696.04 to 257.09, df = 1).

For the second study, a keyword and collocation analysis were completed on the study corpus. Keyness refers to the "aboutness" of a text—that is, what distinguishes a text from other

texts (Egbert & Biber, 2019). The study and reference corpus from study one were used for the analysis. Our research questions were:

1. What are the keywords of online posts about health anxiety?
2. What words distinguish general online posts from online posts about health anxiety?
3. What are the most common collocations of the strongest keyword of online posts about health anxiety?
4. What are the most common collocations of the term "health anxiety" in online posts about health anxiety?

The log-likelihood scale was used to determine significance with $p < .01$ for the top 100 keywords in both corpora. All keywords for study corpus exceeded the critical value (6.63) for significance (G^2 = 9127.9 to 94.26, df = 1, $p < .01$). All keywords for the reference corpus exceeded the critical value for significance (G^2 = 6901.08 to 295.55, df = 1, $p < .01$). Hardie's (2014) log-ratio (*LR)* was used to determine effect size (*LR* = 13.381 to 9.1715). The results from the study corpus were used to identify collocates associated with the top 5 keywords and the term "health anxiety." The mutual information (*MI*) score was used to measure the strength of association between two words of interest. MI scores above three are considered of linguistic interest (Hunston, 2002). The results for most common collocates measures above and below the threshold of three for linguistic interest (*MI* = 11.06 to 2.31).

The results of both studies indicate a significant difference between health anxiety communication and other web-based discourse. Summary scores suggest that communication is authentic and has a degree of negative emotion. Negative emotion amongst those with health anxiety is consistent with previous research (Marcus et al., 2008; Mor & Winquist, 2002). Those in the health anxiety corpus used high levels of first-person pronouns indicative of increased self-

focus similar to other pathologies (Marcus et al., 2008; O'Bryan et al., 2017). Percentage of use of first-person pronoun words corresponded with other anxiety groups in an in-person context reported in previous research (Sonnenschein et al., 2018). Additionally, a linguistic profile emerged. Those with health anxiety present as high in authenticity, low in clout, low in tone, low in analytic thinking, high in first-person pronoun usage, negative emotion and biological terminology. Those with health anxiety were unique in words related to medical conditions or diseases, medication and supplements, medical tests, symptom words, body words, and anxiety words.

The findings may inform clinicians regarding the linguistic attributes of those with health anxiety to increase accurate diagnosis and understanding of the experience of those with health anxiety. Counselor-educators should consider integrating discourse analysis in training programs for counselors-in-training to view the experiences of those with health anxiety, especially descriptions of acute episodes which may not be disclosed during clinical sessions. Researchers may use the results as a baseline measurement for future quantitative or qualitative analysis.

Linguistic Characteristics of Health Anxiety from Online Discourse

by
Christopher McBride

A DISSERTATION

submitted to

Oregon State University

in partial fulfillment of
the requirement for the
degree of

Doctor of Philosophy

Presented February 28, 2022

Commencement June, 2022

Doctor of Philosophy dissertation of Christopher McBride, presented on February 28, 2022.

APPROVED:

_____

Major Professor, representing Counseling

_____

Dean of the College of Education

_____

Dean of the Graduate School

I understand that my dissertation will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my dissertation to any reader upon request.

_____

Christopher McBride, Author

ACKNOWLEDGMENTS

It has been a journey in learning to create and complete this dissertation. Through alterations in subjects and global pandemics, the learning experience has been profound. Many supportive individuals deserve credit for their kind mentorship through this journey.

Thank you to Kok-Mun Ng, Ph.D., who has guided me through this process and the program. His mentorship has been invaluable to my professional and personal growth. Dr. Ng provided an excellent balance between supportive encouragement and high-quality standards.

Thank you to Thom Field, Ph.D., for stepping in as a temporary advisor. Dr. Field was encouraging of the direction of this research project. He challenged me, and my thinking. Dr. Field's input was crucial in the completion of this research.

Cass Dykeman, Ph.D., his passion for corpus studies and for inspiring me on this journey. As the informal methodologist for these studies, Dr. Dykeman’s instruction was a voice in the wilderness. His direction, counsel, and seemingly innumerable cache of resources gave me hope and the courage to continue.

Thank you to my committee member Abraham Cazares-Cervantes, Ph.D., for being a great mentor during internship and for guiding direction during committee meetings. Thank you also to Janet Nishihara, Ph.D., for being willing to be my Graduate Council Representative and ever ready to attend my committee meetings.

Thank you to Mickey Becker, Ph.D., and Trevor Earl, Ph.D. We met every two weeks online to provide support to one another. Even after they finished their doctorates, they continued to meet with me. They answered the call during critical junctures in my process and were always there. Thank you for keeping me on track and for being there for me.

Thank you to my mother and father for their unwavering support of my education throughout the years. They were always so excited and encouraging about daring to take bold next steps in personal and professional development.

Finally, thank you to my beautiful wife, Michelle, and my two children Vincent and Sophia. They missed special occasions while I was away in Wilsonville and endured while I was “in the office” working on my dissertation. They never said one negative word about the program or the time away from the family. I will always be grateful for their support.

TABLE OF CONTENTS

Page

TABLE OF CONTENTS (Continued)

LIST OF TABLES

DEDICATION

This dissertation is dedicated to my dearest wife and life partner Michelle and our lost son William. May all those who suffer with health anxiety find treatment, recovery, and peace from this condition.

Chapter 1: General Introduction

Irvin Yalom (2013) encouraged psychotherapists to enter the client's subjective world. Language is the medium through which individuals communicate their internal psychological experiences (Tausczik & Pennebaker, 2010). Language use can provide information about an individual's beliefs, thinking patterns, social relationships, personality, and fears (Pennebaker et al.,2015). The study of language to understand one's internal psychology and meaning-making has a long history within the mental health field. For instance, Sigmund Freud's (1914) analysis of "Freudian slips" or parapraxis was thought to reveal unveiled hidden intentions and meaning from the unconscious. Narrative therapists investigate ways in which language is used to facilitate personal narrative maps that shape an individual's understanding of the meaning of their life experience over time (White & Epston, 1990). Another example of language-based research within counseling and psychotherapy is the study of metaphor. Mental health researchers note that client language in the use of metaphor can add additional insight into one's view of reality (Ronen, 2011). In addition, the study of metaphor and understanding client language are being studied by cognitive behavioral researchers (Mathieson et al., 2015).

Social media discourse has increased as a source for understanding mental health disorders (Shen & Rudzicz, 2017). Kern et al. (2016) asserted that studying behavior and language on social media has value for the psychology community. While concluding that language is full of psychological information regarding social media studies, Kern et al. also concluded:

> The amount of available data is inconceivable – people leave footprints of their moods, behaviors, personality, and experiences. Social media has become a valuable part of social life, and there is much we can learn by collaboratively studying the tracks left behind, which being cautiously optimistic in our application and approaches. (p. 522)

Therefore, analysis of language used in web-based interactions provides researchers with an opportunity to explore the experiences and meaning-making of certain populations. A targeted analysis may add to critical insight that may inform clinical training, practice, and research.

The corpus linguistic toolset allows researchers to analyze small to very large discourse sets to identify linguistic patterns. Brezina (2018) describes corpus linguistics as "a scientific method of language analysis. It requires the analyst to provide empirical evidence in the form of data drawn from language corpora in support of any statement made about language" (p. 2). Researchers working with corpus linguistics typically construct or use preexisting bodies of texts based on specific criteria. The corpora are analyzed using various techniques to identify keyword lists, collocations, N-grams, linguistic trends, concordance, and so forth. These data points may be used to conduct hypothesis testing, answer research questions, and draw conclusions about the texts and their authors.

Corpus linguistics has aided researchers in identifying the linguistic attributes associated with various mental health disorders and personality characteristics. For example, individuals who qualify for a diagnosis of major depression tend to use more first-person pronouns on social media posts (Chung & Pennebaker, 2007). Another study noted that anger and negative emotion words on Twitter are more prevalent for individuals who qualify for a diagnosis of major depression (Park et al., 2012). In addition, differences in language on social media posts between persons with a posttraumatic stress disorder diagnosis and control groups have been observed (Coppersmith et al., 2014). Findings, as described above, are just a few examples of how counseling and mental health researchers are increasing the knowledge base related to language use for individuals with various diagnoses.

There are several factors that may add support for counseling and mental health researchers to utilize tools within computer-assisted language analysis. First, in 2013, Pennebaker asserted that a revolution is underway in researching linguistic attributes of various populations across the social sciences. Pennebaker takes a psychological approach to the study of language suggesting that the words people use are like fingerprints, revealing information related to an individual's identity and social roles. Regarding counseling, LaGue et al. (2019) asserted, "Corpus linguistic tools can not only give insight into the unique worldviews of clients, authors, or social media participants but provide objective data about effective interventions that can contribute to skill development in counsellors-in-training as well" (p. 13). Greaves and Dykeman (2018) researched the linguistic attributes of individuals with nonsuicidal self-injury. They posited that understanding and analyzing language specific to populations of interest may aid counselors with understanding those individuals' inner experience. According to Greaves and Dykeman, language analysis may help counselors and other helping professionals obtain a more comprehensive understanding of pathology. Considering the aforementioned, corpus linguistics tools, techniques, and analyses are appropriate for mental health counseling-related research.

Health Anxiety

Health anxiety has a long history of documentation within the medical field. The existence of health anxiety was first noted in the 1600s (Fischer-Homberger, 1972). Our understanding of health anxiety and its labeling has evolved over time. Once termed as hypochondriasis, the *Diagnostic and Statistical Manual 5* (DSM-5; American Psychiatric Association, 2013) now categorizes health anxiety under the illness anxiety disorder and somatoform disorders. The belief or fear of having a serious illness is an essential component of hypochondriasis (Warwick & Salkovskis, 1990). Health anxiety, often viewed in a wider

spectrum than hypochondriasis, has the key characteristic of the overapproximation of physically experienced symptoms or perceiving ambiguous physical sensations indicative of a possible lethal medical condition that leads to a fear response (Salkovskis & Warwick, 2001).

Researchers have proposed various theories of health anxiety. Among those theories, the cognitive-behavioral model (CBM) provides a comprehensive explanation and is well researched within the field (Taylor & Asmundson, 2004; Warwick & Salkovskis, 1990). The CBM model outlines the overestimation of physical symptoms and sensations by those with health anxiety. According to the CBM model, a cycle of anxiety is created when an internal or external trigger is experienced; the cycle includes illness thoughts, fear and anxiety, physical sensations, overapproximation or misinterpretation of the seriousness of the symptoms, and behaviors designed to reduced distress (Furer et al., 2007; Walker & Furer, 2009). Individuals with health anxiety often engage in a variety of behaviors to reduce the anxiety such as physical exertion, frequent medical check-ups, searching out medical information, or reading medical textbooks; their behaviors will generally maintain the anxiety (Warwick & Salkovskis, 1990). Counseling researchers may benefit from understanding the language used by individuals with health anxiety and how the linguistic attributes may align with the health anxiety cycle as described in the CBM model.

Individuals with health anxiety can be negatively affected by the condition. Muse et al. (2010) found that 78% of individuals with health anxiety experienced disturbing intrusive cognitions and images. Intrusive images were experienced at a rate of 3.77 times per week. These images impacted the respondents as they reported engaging in avoidance, reassurance seeking, distraction, rumination, and checking-behaviors. Individuals with health anxiety may overuse medical services as a reassurance-seeking behavior which creates a significant burden on

the medical system (Fink et al., 2010). Health anxiety is often underdiagnosed for years and can lead to intense suffering sometimes resulting in suicide (Hedman et al., 2011; Tyrer & Tyrer 2018).

Counseling researchers should be actively engaged in research related to health anxiety as its prevalence is increasing. Tyler et al. (2019) observed health anxiety related to symptoms increased from 14.9% in 2006-2008 to 19.9% in 2008-2010. The emergence of cyberchondria as a type of illness anxiety disorder is noted in the literature. Cyberchondria is defined as excessive and repeated web-based searching for health-related information; this searching may increase anxiety, which may increase searching behaviors in a self-reinforcing cycle (Starcevic & Aboujaoude, 2015; Starcevic & Berle, 2013). The emergence of cyberchondria is another factor for counseling researchers to consider as topics worthy of investigation. As stated above, those with health anxiety may frequently seek reassurance as a component of the condition, and the emergence of social media may be a venue worthy of exploration for researchers.

Reddit is an online social media platform where users can create subreddits for topics of interest. Forums exist for current events such as r/politics to r/BreadStapledToTrees as a forum for images of bread stapled to trees. Subreddits exist as topic-specific boards within the Reddit platform. Mental health-related subreddits may be such things as r/Suicidewatch, r/bipolar, r/depression, and r/healthanxiety. Users locate a topic-specific forum according to their interests. They may write original posts, respond to posts, upvote, downvote, and perform other activities associated with platform engagement. The Reddit platform, therefore, provides an extensive opportunity for linguistic analysis of mental health-related topics.

According to Dean (2021), Reddit.com has 430 million active users. In December 2021, it was the seventh most visited website in the United States (Semrush, 2021). Demographic

reports in February 2021 showed that 18-29-year-olds accounted for 36% of Redditors, 30-49-year-olds accounted for 22%, 50-64-year-olds accounted for 10%, and 65 and older accounted for 3% of users; 18 and under were not reported (Tankovska, 2021a). Currently, the largest user base resides in the United States, accounting for 49.32% of Redditors, followed by the United Kingdom at 7.85%, Canada at 7.76%, Australia at 4.34%, and Germany at 3.11% (Tankovska, 2021b). Twenty-three percent of males and 12% of females in the United States were reported to use Reddit; nonbinary was not reported (Tankovska, 2021c). Tankovska (2021d) reported that of adults polled in the United States, 14% of Hispanics, 4% of Black Americans, and 12% of Whites reported using Reddit.com. Other racial and ethnic backgrounds were not reported. Reddit is primarily coded for the English language without a mechanism for translation. However, subreddits have been created for non-English speakers such as r/espanol/, r/Croatia, r/French, r/newsokur/, and so forth.

Reddit can be a platform for redditors to openly communicate concerns and experiences related to mental health (Gaur et al., 2018). Contributing factors include anonymity and ample character limits on posts and replies. In terms of anonymity, users may create “throwaway” or unidentifiable redditor profiles for privacy or to avoid stigmatization (Park & Conway, 2017). The current character limit is 40,000 per comment or reply. De Choudhury and De (2014) researched mental health disclosures on Reddit and concluded that redditors communicate about significant impacts on their lives. Considering these factors, Reddit presents a rich environment for counseling and mental health researchers to study the communications of individuals with a variety of conditions.

The study corpus for the present dissertation project was created from replies and comments from r/HealthAnxiety for the calendar year 2019. The r/Healthanxiety subreddit was

created in 2012 with English as the primary language. Specific characteristics related to the population for members of this subreddit are not known. The organizers of this subreddit describe the community as “a place for people with Health Anxiety / Illness Anxiety / Hypochondria to come together and start taking control of their disorder” (r/Health Anxiety, 2021). This subreddit lists a membership of 36,500 members as of March 2021. There are eight moderators who monitor the forum and may remove posts that violate forum rules.

In summary, language is a vital part of the human experience that allows for the expression of internal thoughts and processes, and it may reveal attentional focus. As the role of language in human mental health has also been recognized by counseling theorists and practitioners, counseling researchers ought to consider linguistic patterns and interactions as fertile grounds to observe the ever-changing landscape of meaning-making for individuals in relation to understanding and treating mental health issues. Corpus linguistic tools provide a rich set of techniques and methods to investigate significant trends within large bodies of texts allowing researchers to determine both significance and effect size. As health anxiety levels are increasing, perhaps influenced by the increased availability of health-related topics and information widely available on-demand online, the need for mental health counselors and educators to increase their understanding and ability to help address concerns arising from such anxiety appears increasingly urgent. Online mediums and targeted forums present specific and unique locations for mental health researchers to contribute knowledge on health anxiety by making observations of and investigating the development of meaning for individuals who suffer from anxious thinking related to their physiological health.

Dissertation Overview

This dual-manuscript dissertation focuses on identifying the linguistic characteristics of a health anxiety online forum on the platform Reddit. Fundamentally, this research is exploratory to identify the linguistic attributes of discourse within a specific mental health-related forum. Software applications were utilized to identify and analyze core linguistic features related to the uniqueness of this population. The two studies in this dissertation complement one another by highlighting the linguistic qualities of the corpus using a different set of corpus linguistic strategies.

Manuscript 1

The first study investigated the linguistic characteristics of the subreddit r/HealthAnxiety by utilizing the application Linguistic Inquiry Word Count (LIWC). The major sections of LIWC include the linguistic dimensions, other grammar, psychological processes, and summary language variables (Pennebaker et al., 2015). Within the linguistic section, there are 15 subcategories such as pronouns, articles, prepositions, and auxiliary verbs. The other grammar section includes six subcategories such as common verbs, common adjectives, and comparisons. The psychological processes section includes 54 categories such as affective process, positive emotion, negative, and emotion. Summary language variables includes six categories such as authenticity, clout, and emotional tone.

A study corpus was created from a one-year sample of comments and responses from the subreddit r/HealthAnxiety were compiled to comprise the study corpus. The study corpus was compared to the blogs section from the Corpus of Contemporary American English (COCA; The COCA Corpus, 2021). The LIWC application was used to analyze the study corpus, and the results were compared to the reference corpus. The guiding research questions were as follows:

1. What is the score of summary variables about health anxiety?
2. What is the level of use of linguistic processes in online posts about health anxiety?
3. What is the pattern of use of linguistic processes variables in online posts about health anxiety compared to a reference corpus?
4. What is the level of use of psychological processes in online posts about health anxiety?
5. What is the pattern of use of psychological processes in online posts about health anxiety compared to a reference corpus?

The LIWC application was used to analyze the corpus for the variables identified within the literature as being specific to anxiety discourse on Reddit.com (Shen & Rudzicz, 2017). Their variables included authenticity, clout, analytic thinking, first-person singular, anxiety, negative, positive, affect, and feel. Also, these additional variables were added to increase the understanding of the linguistic attributes of the study corpus, first-person plural, second-person singular, third-person singular, and third-person plural. In addition, health related categories within LIWC were added to include the biological processes such as body, health, sexual, and ingestion. Descriptive and inferential statistics were reported.

In reporting inferential statistics, assumptions of normative distribution within texts can be challenging. Language is often differentiated in various contexts. The principle of Ziph's law is often applied to corpus linguistics (Brezina, 2018). According to the distribution of Ziph's law, the most frequent word occurs approximately 7% of the time. In the English language, this is typically the word "the." The second most used word will occur at a rate of one-half of the first used term. The third most used word will occur at a rate of one-third of the most used term, and so on.

As the Ziph's distribution does not adhere to normal distribution, significance and effect size may violate assumptions for parametric tests. Therefore, nonparametric tests were used in analysis. The log-likelihood is a common statistical analysis used within corpus linguistic methodologies (Brezina, 2018). The log-likelihood scale, also referred to as G^2, is similar to the chi-square in that it compares expected values to observed values (Brezina, 2018). For effect size, Bayes Factor (BIC) was used. Results are reported and discussed at the conclusion of the study.

Manuscript 2

Study 2 was a study on keyness and a collocate analysis of the subreddit r/HealthAnxiety. In this study, AntConc (Anthony, 2020), which can identify keywords, collocates, clusters/N-grams, and concordance, among other linguistic attributes of a corpus, was used. The guiding research questions for this study were in reference to the health anxiety corpus:

1. What are the keywords of online posts about health anxiety?
2. What words distinguish general online posts from online posts about health anxiety?
3. What are the most common collocations of the strongest keyword of online posts about health anxiety?
4. What are the most common collocations of the term "health anxiety" in online posts about health anxiety?

The most frequent unique words in a corpus are indicative of the aboutness, sometimes referred to as the keyness of the corpus (Brezina, 2018). The corpus of reference was the COCA blog corpus (The COCA Corpus, 2021). The top 100 keywords, compared to the reference corpus, are reported along with their associative keyness score. Hardie's (2014) log-ratio test reports the effect size of the keywords listed.

For collocations, the top five keywords and the term “health anxiety” from the study corpus were selected as node words and evaluated. Regarding collocations, Brezina (2018) wrote, “Collocations are combinations of words that habitually co-occur in texts and corpora. Collocations can be based either on frequency alone or, as is more common, on a statistical measure” (p. 66). Identification of collocates can assist in providing further information related to the node word under investigation. The mutual information test, which measures the strength of the relationship between variables, was used.

Summary

I have in this chapter articulated the central foci of the two corpus linguistic studies on health anxiety based on Reddit and described their thematic relationships. Language is central to the human experience and impacts understanding and meaning creation. As online forums and social media platforms increase in prevalence, the emerging discourse is worthy of scholarly investigation. Health anxiety has increased over the past decade, which may be impacted by online behaviors. Understanding the unique language of support forums may benefit the mental health community and could be a starting point for further research through an interdisciplinary lens. Chapter 2 is a study using the LIWC application to identify and compare the linguistic attributes of the study corpus and the reference corpus. Chapter 3 is a keyword and collocation analysis of corpus constructed from the r/healthanxiety subreddit. Finally, in Chapter 4, I summarize major findings from both studies and synthesize the findings. Chapter 4 includes research, training, clinical recommendations, and conclusion.

Description of Terms

AntConc – A software application (Anthony, 2020) designed to analyze corpora. Output may include keyword lists, word lists, collocates, N-grams, and concordance.

Collocations – A collocate, or collocation, refers to the association of words that appear in proximity to one another (Weisser, 2016).

Corpus (pl. corpora) – A corpus is a collection of texts gathered based on criteria for linguistic analysis. These texts may originate from the spoken or written word or a combination of both (Weisser, 2016).

Health anxiety – Health anxiety, is the over approximation of physically experienced symptoms or perceiving ambiguous physical sensations indicative of a possible lethal medical condition that leads to a fear response (Salkovskis & Warwick, 2001).

Health anxiety cycle – A cycle based on the cognitive behavioral model of health anxiety. Internal or external triggers lead to an illness thought and anxiety and fear (Furer et al., 2007). The anxiety leads to bodily sensations and the interpretation of sensations as a serious illness. Subsequently, this leads to coping behaviors including reassurance seeking, checking, safety signals, or avoidance. The coping behaviors lead to a temporary reduction of anxiety and fear.

Keywords – A word that is more frequent in a corpus understudy than in a comparison corpus (Hardie et al., 2006).

Keyness – Keyness, derived from keywords, refers to words that may be important in determining the uniqueness or aboutness of a corpus (Phillips, 1989).

Linguistic Inquiry Word Count (LIWC) **–** A software application designed to designate words into categories based on frequency of use within a corpus (Pennebaker et al., 2015).

Linguistic process – A set of categories with the LIWC dictionary (Pennebaker et al., 2015). Linguistic process may include first-person pronouns, third-person pronouns, prepositions, and so forth.

Node word – A node word is the central word or grammatical structure that serves as the basis for a search for collocations or concordance (Weisser, 2016).

Psycholinguistic processes – A set of categories within the LIWC dictionary (Pennebaker et al., 2015). The psycholinguistic process may include positive emotions, negative emotions, cognitive processes, and so forth.

Register – A classification of texts that may include purpose of the text, intended audience, narration or description, formality, and so forth (McEnery & Hardie, 2011). May be synonymous with *genre,* although some scholars may differentiate these terms.

Stop words – Function words that may add "noise" to a corpus when determining keywords and collocations—words such as "the," "as," "of," and "also."

Token – A single occurrence of a word within a corpus.

Type – A unique word form within a corpus.

Chapter 2

Linguistic Characteristics of Health Anxiety Posts on Reddit

Christopher McBride

Kok-Mun Ng

Thom Field

Authors note

Christopher McBride, Counseling Academic Unit, Oregon State University; Kok-Mun Ng, Counseling Academic Unit, Oregon State University; Thomas Field, Counseling Academic Unit, Oregon State University.

The research contained in this manuscript was part of the first author's dissertation research project.

Correspondence concerning this article should be addressed to Christopher McBride, Email: mcbrichr@oregonstateu.edu

Abstract

Computer-assisted linguistic analysis is providing information for counseling and other helping researchers to increase their understanding of the internal experiences of individuals with varying pathologies. Health anxiety is a condition that is increasing within the larger population. Social media use may be influencing the evolution of symptoms, as well as giving rise to new conditions such as cyberchondria. Little is known about the linguistic qualities of health anxiety within social media forums. The purpose of this study was to evaluate the linguistic characteristics of users on a health anxiety forum hosted on Reddit. We constructed a study corpus from a one-year sample of posts and comments from the subreddit r/Healthanxiety. We compared the study corpus to a reference corpus created from the blogs section contained within the Corpus of Contemporary American English (COCA). We used the Linguistic Inquiry Word Count (LIWC) application to identify summary, linguistic process, and psychological process variables. Log-likelihood (G^2) and Bayes factor (BIC) analysis showed significant differences between the study and references corpora. We discuss implications of the findings to previous research and implications for the CBT model of the health anxiety cycle. We conclude with recommendations for counselor educators, clinicians, and researchers.

Keywords: health anxiety, r/Healthanxiety, corpus linguistics, LIWC, Reddit, linguistic attributes

Linguistic Characteristics of Health Anxiety Posts on Reddit

Linguistic attributes related to mental health disorders, such as illness anxiety disorder, are in the early phases of examination. According to Tyrer et al., (2019) the prevalence of health anxiety is increasing. Health anxiety is often underdiagnosed and those with health anxiety may suffer intense distress for years (Hedman et al., 2011; Tyer & Tyer, 2018). Health anxiety is a significant feature of several mental health disorders that first emerged in medical literature in the 1600s (Fischer-Homberger, 1972). Previously known as hypochondriasis, health anxiety consists of negative interpretations and fears related to either ordinary or unusual bodily sensations and observations (Salkovskis & Warwick, 2001).

Health-related anxieties are categorized within the somatic and illness anxiety disorders in the fifth edition of the *Diagnostic and Statistical Manual of Mental Disorders* (DSM-5; American Psychiatric Association, 2013). Health anxiety may be useful for help-seeking behaviors; however, elevated levels of health anxiety may cause significant distress (Asmundson & Taylor, 2020). Health-related anxiety is observed on a continuum from mild to extreme symptoms (American Psychiatric Association, 2013). In more severe cases, some of those with health anxiety experience fear and preoccupation that they suffer from a serious, often fatal condition without a medical condition to support such fears (Warwick & Salkovskis, 1990). Sometimes, health anxiety leads a suicide (Tyrer & Tyrer, 2018).

Health anxiety is often comorbid with other anxiety-related disorders such as generalized anxiety disorder, a specific phobia, and panic disorder or mood disorders such as major depression (Abramowitz et al., 2007; American Psychiatric Association, 2013). Health anxiety may develop in the context of a predisposition for anxiety, distressing experiences with death, serious illness, or other stressful life events (Walker & Furer, 2008).

Multiple models are used to conceptualize health anxiety. However, the cognitive-behavioral model (CBM) of health anxiety is well researched and offers a comprehensive explanation of the condition (Taylor & Asmundson, 2004; Warwick & Salkovskis, 1990). Based on CBM, Furer et al. (2007) described a cycle of health anxiety in which internal or external triggers set off a cascade of cognitive, affective, behavioral, and physiological responses. Internal or external triggers are conceptualized as the entry point to the cycle. Internal triggers may consist of thoughts or memories related to health concerns or bodily symptoms or alterations of physical functioning. External triggers may consist of exposure to health-related stimuli that evokes an anxiety response such as reading a health-related news article or hearing about the illness of another person. Internal or external triggers elicit illness thoughts, anxiety, and fear. Increased bodily sensations follow the anxiety and fear. The bodily sensations are interpreted as signs of a serious illness. Subsequently, the individual will engage in either reassurance-seeking behaviors or avoidance techniques to self-soothe. Reassurance behaviors include increased checking, asking family members for reassurance, increased medical appointments, and other actions. In terms of impact of health anxiety, Fink et al. (2010) forward that those with health anxiety may overuse medical services for reassurance, creating a significant burden on the medical system.

Further research should be conducted on health anxiety because the prevalence of the condition is increasing (Tyrer et al., 2019). Tyrer et al. reported an increase in health anxiety prevalence from 14.9% in 2006-2008 to 19.9% in 2008-2010. Social media and internet usage may be a possible explanation. For example, cyberchondria is a type of illness anxiety, originating from web-based behaviors. Cyberchondria is defined as repeated or excessive web-based searching for health-related information, which increases anxiety (Starcevic &

Aboujaoude, 2015; Starcevic & Berle, 2013). Individuals seek to lessen concerns related to medical conditions but often find new information that increases rumination and anxious feelings associated with a current or discovered malady via online research.

In relation to noting an increase in the prevalence of health anxiety, the emergence of the COVID-19 pandemic has had an impact on social media interactions. While comparing 60 million Tweets from Twitter.com from March 24, 2020, to May 24, 2020, to 40 million Tweets from the previous year, Saha et al. (2020) found that negative mental health expressions on social media increased approximately 14%, while support expressions increased approximately 5%. Saha et al. noted that language appeared to reduce and normalize over time, suggestive of the population's ability to adapt to the present circumstances. Holmes et al. (2020) set forth a research priority list for mental health researchers to investigate the impact of the COVID-19 pandemic, including through media and social media interactions at the population level. The present study may provide a baseline from which results can be compared for future research.

Linguistic-based research can be useful to clinicians. Pennebaker et al. (2003) asserted that words individuals use can be informative of their social, mental, and even physical state. This is not a new idea in the study of human development and psychology. Freud (1901) documented the presence of parapraxes in which slips of the tongue may reveal deeper information from the individual. Developing an understanding of language and how language is used among certain populations may provide information relevant to the personal experiences of those with mental health conditions (Greaves & Dykeman, 2018). For example, Greaves and Dykeman observed high frequencies of the terms “feel,” “help,” and “scar” when studying the language of nonsuicidal self-injury (NSSI) posts on Reddit (p. 11). Understanding that individuals who struggle with NSSI may frequently discuss their scars provides counselors

additional information and insight into the experiences of those with NSSI. Another example is the research by Al-Mosaiwi and Johnstone (2018) who observed an increased frequency in absolutist words across social media posts specific to anxiety, depression, suicidal ideation disorders, borderline personality disorders, and eating disorders. They theorized that this finding adds to the empirical evidence for clinicians who want to increase cognitive flexibility and diminish absolutist cognitions.

For clinicians, obtaining information related to linguistic markers across various populations may be useful for adequate detection, diagnosis, and treatment. De Choundry et al. (2013), in their analysis of the linguistics of postpartum depression, asserted that developing tools for automated analysis to monitor social media posts may provide health care providers the ability to identify markers indicative of increased depression levels. Identifying such markers can lead to increased efficiency and scope of possible interventions.

Corpus linguistic studies that leverage advancements in computer technologies to analyze larger portions of language are informing scholars across multiple disciplines (Pennebaker, 2012). A corpus is a collection of spoken or written text based on specific criterion that are used for linguistic analysis (Weisser, 2016). A corpus can be constructed via a variety of criteria. For example, a researcher may be interested in contemporary research. As such, they build a corpus using a synchronic design. Further, a corpus can be static, meaning the size is fixed (not intended to be augmented or diminished) or dynamic (expected to change over time to reflect the ever-evolving nature of language; Weisser, 2016, p. 24). Researchers interested in historical or comparative studies use a diachronic design which allows for tracking linguistic changes over time (Brezina, 2018).

Corpus linguistic techniques have evolved. Initially, corpus linguistic studies relied on a manual method of analysis. The manual process increased time and cost and introduced additional opportunities for error. However, the emergence of computer-assisted research allows for the analysis of large sets of texts, which have increased flexible and efficient research tools. One such linguistic application is the Language Inquiry and Word Count (LIWC) software application (Pennebaker et al., 2015). LIWC was first developed in the early 1990s with multiple iterations; the most recent version of LIWC emerged in 2015 (Pennebaker et al., 2015). The function of LIWC is to report a percentage of the type of word usage based within a corpus and summary variables calculations.

The current iteration of LIWC utilizes the 2015 LIWC dictionary. The LIWC2015 dictionary is comprised of 90 output variables and 6,400 words, specific emoticons, and word stems (Pennebaker et al., 2015). Output variables consist of global variables such as affective process. More specific variables may be listed under a global variable category, such as positive emotion, which exists under the global category of affective process. Words may be included under multiple categories. For example, using the word “cried” may increase the frequency of the percentage of usage within the text under the overall effect, as well as negative emotion, sadness, verbs, and past focus. The LIWC application is not coded to be able to differentiate between contexts within a language. For example, LIWC output does not distinguish between a text that uses sarcasm versus sincere expression. However, output derived from analysis may indicate that text scores low in the authentic summary variable.

The LIWC application has three broad classification categories, including summary process variables, linguistic process variables, and psychological process variables (Pennebaker et al., 2015). Summary variables provide information related to analytic, authentic, clout, and

emotion tone of the text analyzed. The linguistic process category includes variables such as pronouns, articles, prepositions, auxiliary verbs, common adverbs, conjunctions, and negations. Psychological processes category includes variables that identify affective, social, cognitive, perceptual, biological, drive, time orientation, relativity, personal concerns, and informal language words.

Researchers have utilized LIWC to study mental health, personality, and positive aspects of social media use by analyzing social media postings. For example, Coppersmith et al. (2014) identified language differences in posttraumatic stress disorder, social anxiety disorder, and bipolar disorder by analyzing user posts on a social media platform using LIWC. Park et al. (2013) identified linguistic markers to distinguish Tweets from users with posttraumatic stress disorder versus other disorders such as social anxiety disorder, bipolar, and unipolar depression based on the linguistic characteristics. In addition, researchers have utilized LIWC to study personality factors and to draw conclusions about the effectiveness of online forum participation over time (see Park & Conway, 2017; Qui et al., 201).

Reddit presents a favorable environment to observe mental health dialog among social media users. Reddit users may remain anonymous and create “throwaway” or unidentifiable accounts to candidly discuss information that redditors want to keep private due to stigmatization or other reasons (Park & Conway, 2017). Reddit offers redditors a platform to openly discuss mental health concerns due to anonymity and large character limits in posts and comments (i.e., 40,000-word character limit per post) to describe their experiences (Gaur et al., 2018). De Choudhury and De (2014) investigated mental health disclosures on Reddit and concluded that users write about the significant impact on their lives.

Reddit.com is an online social media platform that has 430 million active users (Dean, 2021). As of December 2021, it was the seventh most visited website in the United States (Semrush, 2021). On Reddit, users are termed "redditors," and they join communities and subreddits, submit comments and replies, upvote comments or replies, downvote comments or replies, and engage in various other ways with the site content. As of February 2021, 18-29-year-olds comprised 36% of the redditors, 30-49-year-olds comprised 22%, 50-64-year-olds comprised 10%, and 65 and older comprised 3% (Tankovska, 2021a). The United States comprises 49.32% of Reddit users, followed by the United Kingdom at 7.85%, Canada at 7.76%, Australia at 4.34%, and Germany at 3.11% (Tankovska, 2021b). Within the United States, 23% of males and 12% of females reported using Reddit; nonbinary genders were not reported (Tankovska, 2021c). Tankovska (2021d) reported that of U.S. residents, 14% of Hispanics, 4% of Black Americans, and 12% of White Americans use Reddit.com; other racial and ethnic groups were not reported. Reddit is written in the English language without a mechanism to change the language or alternate language. However, subreddits are available for non-English speakers such as r/espanol/, r/Croatia, r/French, and r/newsokur/.

A health anxiety-related community on Reddit is the subreddit r/HealthAnxiety. Per the r/HealthAnxiety description, this subreddit was created in 2012 (r/Health Anxiety, 2021). The description of the purpose of r/HealthAnxiety says it is "a place for people with Health Anxiety/Illness Anxiety/Hypochondria to come together and start taking control of their disorder" (Health Anxiety, 2022). As of January 2022, this subreddit reported 52,300 members with eight listed moderators. As anyone on the Internet may become a redditor, the nationalities within the subreddit r/HealthAnxiety are not known at this time.

Although Reddit represents one of the largest and most diverse social media websites, limited research has been conducted investigating its linguistic characteristics and health anxiety. Among the limited studies related to linguistic characteristics on Reddit, Shen and Rudzicz (2014) constructed a corpus analysis using Reddit to study anxiety-related posts. The researchers utilized LIWC to determine common linguistic frequencies amongst words, including anxiety, negative, dictionary words, affect, first-person singular, emotional tone, authentic, clout, analytic, and feel. These researchers called for more research to be conducted to further understanding and detection-related tools for diagnosis.

A reference corpus should be considered a source that provides general information about the language (Leech, 2002). Leech asserts that the reference corpus should be draw from multiple sources to ensure diversity, that it should be considered a standard among the user community, and that it should be used as a comparison against some other variety of language. The Corpus of Contemporary American English is currently one of the most widely used corpora (COCA; The COCA, 2001). The COCA blogs section includes blogs postings as identified by Google.com from academic, argument, fiction, info, instruction, legal, news, personal, promotion, and review web pages. Considering the aforementioned, the COCA blogs section is a diverse set of online blog postings which is widely accepted as the standard by which other English corpora may be compared.

To date, to our knowledge, research investigating the linguistic characteristics specific to health anxiety from the subreddit r/HealthAnxiety has not been conducted. Our study sought to fill a research gap. The purpose of this study was to gain descriptive information about linguistic characteristics of health anxiety from posts and replies within the subreddit r/HealthAnxiety and to compare the data to a reference corpus. The research questions (RQ) were:

1. What is the score of summary variables about health anxiety?
2. What is the level of use of linguistic processes in online posts about health anxiety?
3. What is the pattern of use of linguistic processes variables in online posts about health anxiety compared to a reference corpus?
4. What is the level of use of psychological processes in online posts about health anxiety?
5. What is the pattern of use of psychological processes in online posts about health anxiety compared to a reference corpus?

Method

Design

In this study, a synchronic corpus linguistic design was used to explore the linguistic characteristics of health anxiety posts on Reddit (Weisser, 2017). The reference corpus was the blog corpus extracted from COCA (The COCA Corpus, 2021).

Power Analysis

To test the power of the results for the research questions, an a priori power analysis was completed using G*Power (Faul et al., 2009). Log-likelihood is a derivation of the $\chi 2$ test. The appropriate effect size is Cohen's *w* which was assumed at a medium effect of .3 (Rosnow & Rosenthal, 2003). The input parameters were: (a) test-family- $\chi 2$ test; (b) statistical test- goodness-of-fit tests: contingency tables; (c) type of power analysis – a priori: compute required sample size- given, power, and effect size; (d) $w = 0.3$; (e) power (1- β error probability) = 0.90; (f) $\alpha = .05$; and (g) degrees of freedom (*Df*) = 1. The G*Power output indicated a sample size of 117 for an actual power of 0.90.

Study Corpus

Register, Scope, and Sources

The register is internet-based informal discourse. The subregister is health anxiety communication. The scope is comments and replies on a public forum dedicated to health anxiety. The source is posts and comments from the subreddit r/HealthAnxiety for the calendar year 2019. This year was selected as it contained the most recent data prior to the COVID-19 pandemic. COVID-19 may have had an impact on health-related communication and thus skewed the results of normative communication on a health anxiety discussion board. A full calendar year is viewed as optimal as it controls for possible seasonal changes in communication.

Preprocessing

The corpus was extracted using the API interface within the Reddit database. The corpus was comprised of posts and comments written in monolingual English. To protect users' privacy, the corpus excluded usernames, location, personalized URLs, hashtags, or location of posts. The corpus data were preprocessed to ensure the greatest opportunity for LIWC to capture all word frequency counts and variable designation. Non-U.S. dialect spelling was converted to U.S. dialect (e.g., “colour” UK spelling to “color” U.S. spelling). Finally, linguistic variants and out-of-vocabulary words and expressions were converted for lexical normalization when possible. The final corpus was formatted into text file format for LIWC processing.

Reference Corpus

Register, Scope, and Sources

The register is internet-based informal discourse. The scope is blog posts as identified from Google. The source is the blog corpus from the COCA (The COCA Corpus, 2021). The blog section contained online articles for the year 2012.

Preprocessing

A similar preprocessing method was applied to the reference corpus as completed with the study corpus. The blog corpus selected was from the blogs section of the COCA. Non-U.S. dialect words were converted to U.S. English (e.g., “colour” UK spelling to “color” U.S. spelling). Linguistic variants and out-of-vocabulary words and expressions were converted where possible (e.g., “LOL” to “laugh out loud”). The final corpus was output into a .txt format. Due to the size of the reference corpus, it was divided into six separate text files for processing by the LIWC application.

Instrumentation

Linguistic Inquiry and Word Count (LIWC)

This study used LIWC (Pennebaker et al., 2015) to analyze results from the corpus of interest. Variables for the study were those identified by Shen and Rudzicz (2017) as pertinent to anxiety posts on Reddit. The variables included authentic, clout, dictionary words, emotional tone, analytic thinking, first-person singular, anxiety, negative, affect, and feel. Linguistic variables of first-person plural, second-person singular, third-person singular, and third-person plural were included in the analysis. In addition, this study included the variables from the biological subcategory—body, health, sexual, and ingestion. These categories are associated with physical sensations and are assumed to be of interest to those with health anxiety.

Data Analysis

Data analysis consisted of descriptive statistics and corpus comparison to identify differences between the variables using inferential methods. A total word count and percentage of the corpus were reported for all identified variables within the study. The LIWC data were separated by specific variables for summary, linguistic, and psychological process variables.

Utilizing this principle, the LIWC output for the corpus data were analyzed by applying the log-likelihood (G^2) and Bayes factor (BIC) tests. Log-likelihood and Bayes Factor (BIC) was calculated using the *R* statistical analysis application. The *p*-value was assumed at less than .01 ($p < .01$) with a critical value for significance ($G^2 > 6.63$). The log-likelihood formula requires that data from the two sources are independent (Brezina, 2018). Sources from the reference corpus, the COCA, and posts and comments from Reddit were not present. In LIWC, variables other than the summary categories are reported as percentage of the text. As such, percentages were converted to raw frequencies for log-likelihood and Bayes factor analysis. The study corpus was compared to the reference corpus to derive the log-likelihood calculation. Bayes information criterion (BIC) was used to determine evidence against $H_{0,}$ that there is no difference in frequencies for linguistic variables between the study and the reference corpus.

Summary variables were provided as a standardized score of the category within the corpus. The standardized score is derived from a proprietary calculation within LIWC. As such, raw frequencies could not be derived from the information provided, and inferential statistical analysis was not performed for summary categories. Summary scores are reported using the standardized score from the LIWC output.

Results

Posts and comments were extracted using the Reddit API for the year 2019 and combined into a single text file. Preprocessing procedures were completed on both the study and reference corpus. In corpus linguistics nomenclature, a single occurrence of a word within a corpus is referred to as a "token" and unique word form is referred to as a "type" (Brezina, 2018).

After the preprocessing, the study corpus contained 5,461,459 tokens. The total posts and comments accounted for 80,129 total entries for an average of 68.15 tokens per entry. After

preprocessing, the reference corpus had 106,568,862 tokens for 98,796 entries for an average of 1,078.67 tokens per entry.

In LIWC, output for variables is reported as a percentage of the text, except for summary variables. Percentages were converted to frequency count for log-likelihood and Bayes Factor (BIC) analysis. Due to the size of the reference corpus and limitations with LIWC in processing larger bodies of text, the reference corpus was divided into six separate files for analysis. LIWC percentages were then summed and divided to derive the mean for the categories under study. The mean of the category was used to obtain the actual word count.

For descriptive statistics, percentage of text is reported in Table 2.1. Percentage of overall text was calculated into actual word count. For RQ1, the broad summary scores about health anxiety from the LIWC application are listed in Table 2.1. Results for RQ2 and RQ3 are also presented in Table 2.1.

Table 2.1

LIWC Descriptive Statistics (RQs 1-3)

Variable Category	Variable	Study Corpus		Reference Corpus		Summary Var	
		Raw Ct	% of All Words	Raw Ct	% of All Words	Study Corpus Stand. Score	Ref. Corpus Stand. Score
Broad	Analytic					31.67	78.58
Broad	Clout					24.10	62.3
Broad	Authentic					85.48	32.36
Broad	Tone					7.66	48.26
Ling	1ps	442924.3	8.11	2360500	2.15		
Ling	1pp	4915.31	0.09	383647.9	0.36		
Ling	2p	121243.4	2.22	1445784	1.35		
Ling	3ps	20206.4	0.37	1152720	1.08		
Psych	Anxiety	104860	1.92	282407.5	0.26		
Psych	Affect	385032.9	7.05	5264502	4.94		
Psych	P Emo	146913.2	2.69	3268112	3.06		
Psych	N Emo	232112	4.25	1916463	1.79		
Psych	Anger	27307.3	0.50	637637	0.59		
Psych	Sad	232112.86	0.47	1916463.8	0.32		

Bio	Body	131075	2.40	458246.1	0.43
Bio	Health	198251	3.63	619875.5	0.58
Bio	Sexual	9284.48	0.17	150972.6	0.14
Bio	Ingest	36591.78	0.67	396080.9	0.37

Note. Raw Ct is the frequency of the word occurrence in corpus, Stand. Score is Standardized score.

For RQ4 through RQ6, Table 2.2 details the inferential statistics for the results. Log-likelihood (G^2) shows significance across variables in either over or underuse. All variables exceeded the critical cut-off of 6.63 for $p < .01$ for significance (G^2 = 484579.4 to 276.0733, df = 1, $p < .01$). Bayes Factor (BIC) scores over 10 indicate very strong evidence against the null hypothesis (Wilson, 2013). Bayes Factor (BIC) scores indicate very strong evidence across all variables (BIC = 465696.04 to 257.09, df = 1). Overuse variables in the study corpus were 1st person singular, 2nd person singular, anxiety, negative, affect, negative emotion, sad, body, health, sexual, and ingest. Underuse in the study corpus was 1st person plural, 3rd person singular, positive emotion, and anger.

Table 2.2

LIWC Inferential Results (RQs 4-5)

Category	**Process**	G^2	**Overuse/ Underuse**	BIC	BIC **Descriptor**
1st sing	Linguistic	484579.4**	+	465696.04	Very Strong
1st plur	Linguistic	15355.87**	-	15300.14	Very Strong
2nd sing	Linguistic	24064.74**	+	23651.19	Very Strong
3rd sing	Linguistic	33384.72**	-	33103.77	Very Strong
Anxiety	Psychological	210893.3**	+	209439.68	Very Strong
Affect	Psychological	43479.02**	+	41007.63	Very Strong
Po. Emo	Psychological	2589.557**	-	2495.14	Very Strong
Neg. Emo	Psychological	125967.6**	+	122739.83	Very Strong
Anger	Psychological	897.2782**	-	873.79	Very Strong
Sad	Psychological	2929.07**	+	2910.54	Very Strong
Body	Psychological	215191.4**	+	213133.85	Very Strong
Health	Psychological	358659.3**	+	353732.31	Very Strong

Category	**Process**	G^2	**Overuse/ Underuse**	BIC	BIC **Descriptor**
Sexual	Psychological	276.0733**	+	257.09	Very Strong
Ingest	Psychological	9955.97**	+	9890.19	Very Strong

Note: The G^2 critical value for $^{**}p < .01$ is 6.63. $df = 1$. G^2 is Log-likelihood. Overuse/Underuse represents comparison of study corpus versus reference corpus. Overuse (+) means more frequent is study corpus. Overuse (-) means less frequent in study corpus. BIC is Bayes Information Criteria. Scores over 10 indicate very strong evidence against H_0.

Discussion

An important task for counselors is to enter the subjective world of their client (Yalom, 2013). The method of corpus linguistic studies allows for mental health researchers to enter the world of certain pathologies at a group level, observing the various linguistic traits and patterns of communication. Such an endeavor can assist counselors in understanding the experiences of individuals to provide a view into the internal experiences that may not normally be observed within the counseling room.

Specifically, the purpose of this study was to gain a better understanding of the linguistic attributes of health anxiety communication on Reddit. That is, we were interested in how communication transpires for this subreddit dedicated to health anxiety. The results painted an interesting picture of the community. Primarily, the findings show significant differences between the study corpus and the reference corpus. At a high level, the key findings demonstrate that communication is authentic with a high degree of self-focus laced with negative emotion and anxiety words.

In considering the summary variables, the results indicated that reddit users are authentic when communicating. This finding supports previous research (De Choudhury & De, 2014; Gaur et al., 2018). De Choudhury and De found that posts and comments on mental health-related subreddits can have a high degree of self-disclosure due to anonymity and throwaway accounts,

among other factors. Considering the authenticity score (85.48), the health anxiety corpus suggests that posts may reliably reflect the experiences of posters on the board.

The most significant difference in linguistic variables was the first-person pronoun (FPP) usage in the study corpus. This corresponds with the findings of Sonnenschein et al. (2018). Sonnenschein et al. (2018) compared LIWC scores for FPP across mood and anxiety disorders in a clinical context. For anxiety disorders without comorbid depression, the LIWC score was 8.10%. This aligns closely with the study corpus result of FPP of 8.11%. This corollary may be indicative of pronoun usage being consistent across contexts such as online versus in-person communication. Self-focus is defined as "an awareness of self-referent, internally generated information that stands in contrast to an awareness of externally generated information derived through sensory receptors" (Ingram, 1990, p. 156). The degree to which someone refers to themselves in writing or verbally can be indicative of self-focus (Tausczik & Pennebaker, 2010). Tausczik and Pennebaker (2010) asserted that when individuals experience emotional or physical pain their attention is drawn inward, which explains increased attention towards the self and elevated use of FPP.

The CBT model provides a possible explanation for heightened self-focus. The CBT model of health anxiety describes a cycle in which an internal or external trigger elicits an illness thought with anxiety and fear (Furer et al., 2007). Subsequently, bodily sensations may be interpreted as a sign of a serious disease or disorder which leads to checking and reassurance seeking. This may temporarily reduce anxiety or fear but reinforces the behavior. The nature of this model relies on self-focus as the individual is monitoring, evaluating, and misinterpreting internal and external queues. As the behavior or reassurance seeking is self-reinforcing, those with health anxiety may become increasingly habituated to the cycle and increase their self-

monitoring for signs of trouble. For example, internal triggers lead to an illness thought which evokes anxiety and fear. Self-focus levels may be high as the individual is attuned to what is transpiring internally and interpreting and misinterpreting internal and external triggers. Further, increased attention to bodily sensations may also reinforce the awareness of internally generated information as fear and increased bodily sensations lead to an overapproximation of the severity of the symptoms experienced. Increased self-focus could be indicative of a cognitive bias that can be closely connected with both acute and chronic negative affect (Mor & Winquist, 2002). This cognitive bias may be an explanation for elevated scores of psycholinguistic variables.

The psycholinguistic variables negative, affect, and negative emotions showed significant difference compared to the reference corpus. These results are not surprising as the *DSM-5* (American Psychiatric Association, 2013) notes significant distress related to symptoms for both illness anxiety disorder and somatoform disorder. Taken with the increased self-focus, these findings agree with Mor and Winquist (2002) regarding increased self-focus and negative affect. Further, the elevated use of negative, affect, and negative emotion words support previous research that found a correlation between health anxiety and negative affect (see Marcus et al., 2008). O'Bryan et al. (2017) found that those with greater heath preoccupations experienced emotions more frequently, intensely, and for longer durations. While the results within this study cannot verify rumination or cognitive distortions, the high degree of body, health, and ingest variables suggest an emphasis on writing about physiological topics. These results may mirror a preoccupation or rumination common among those with health anxiety (Marcus et al., 2008). Interestingly, the results of this study show twice the level of anxiety and sad language use than reported by Sonnenschein et al. (2018) for anxiety without comorbid depression. Whether the

difference is due to context or linguistic variation between health anxiety and general anxiety disorders is unknown.

In terms of the biological variables, body, health, and ingest scores were elevated within the health anxiety corpus. This seems consistent with the purpose of the discussion board and, as mentioned earlier, may also be indicative of rumination regarding the topic on health-related matters. Bodily related matters are essential within the framework of health anxiety as the locus of anxiety is often essentially physiological experiencing and interpretation. Considering the aforementioned, words associated with health-related anxiety may often be physiological in nature. Within the CBT model, internal or external triggers are the impetus of illness thoughts and anxiety sensations (Furer et al., 2007). This may lead to bodily sensations and catastrophizing thoughts regarding a possible illness. Subsequently, reassurance seeking, checking, safety signals, and avoidance occur. Knowledge regarding the stage of the model by which redditors most commonly post is beyond the scope of this research. However, it may be possible that users post for reassurance seeking and checking as item level posts often detail symptoms along with expressions such as experiencing significant health anxiety "at the moment" which redditor will often abbreviate to "atm."

In summary, these findings highlight that posts and comments within the health anxiety corpus are authentic, have a high degree of self-focus, have a high degree of negative affect, and are focused on topics related to the body. The findings may be conceptualized within the CBT model for health anxiety and are consistent with previous research regarding health anxiety. These linguistic characteristics describe how those with health anxiety communicate at a group level. This perspective from a group may assist mental health professionals in garnering a deeper understanding of the experiences of those with health anxiety, especially if they choose to peruse

the posts and comments on the subreddit of study. As Yalom (2013) emphasized the importance of entering the individual experience of clients, review of postings on this subreddit allows for professionals to enter the world of those who are suffering with health anxiety and to potentially obtain a new perspective, especially of those during acute anxiety episodes potentially not seen in a professional setting.

Clinical Implications

The purpose of this study was to determine the linguistic attributes of health anxiety communication from an online forum. The results describe a group that communicates authentically, with high degrees of negative affect and biological words consistent with the pathology and literature on health anxiety. Considering these factors, the current forum appears as a valid representation of the population.

From a counselor education perspective, counselors-in-training may benefit from reading the posts and replies as an addition to their current training to gain further understanding of this condition. At the post level, counselors-in-training may benefit from descriptions of acute health anxiety which are often accompanied with the acronym “atm” which stands for at-the-moment. Also, counselors-in-training could benefit from identifying accompanying cognitive distortions, such as catastrophizing, to increase their understanding of how such distortions interact with negative emotional affect. Likewise, practicing counselors may benefit from reading posts and replies as they increase their competencies when interacting with this population.

At an enterprise-level, there may be opportunities to expand the field of the counseling practice. For example, De Choundry et al. (2013) suggested creating online monitoring of postpartum depression forums for automated interventions based on linguistic markers contained within posts. As more innovations of real-time monitoring are constructed, counselors may find

opportunities contracting with larger platforms such as Reddit to evaluate and offer online community mental health services. If baselines are established across linguistic attributes, elevated linguistic attributes at the comments or replies level may serve as markers for intervention. Online interventions such as internet-based CBT for health anxiety have demonstrated effectiveness in treating health anxiety (Newby et al., 2018). Perhaps such interventions, psychoeducational material, or referral sources could be offered to redditors based on real-time linguistic monitoring. As the counseling profession continues its growth, more counselors may find opportunities working at an enterprise-level for larger hosting institutions.

Finally, there are opportunities for advancement in assessment and evaluation. As linguistic attributes are identified across mental health conditions, it is conceivable that narrative writing may be evaluated to augment assessment data relevant to diagnosis and personality. For example, Holtzman et al. (2019) demonstrated a modest link between LIWC categories and narcissistic personality disorder. As the field continues to develop tools and methods related to linguistic attributes and mental health, opportunities towards new types of assessment and measurement will be possible.

Research Implications

The present study outlined the linguistic attributes of Redditors on a health anxiety subreddit. Building on the LIWC variables identified by Shen and Rudzicz (2017), an increase in understanding the linguistic characteristics for online communication regarding health anxiety communication is provided in this study. Fundamentally, the results of this research present a consistent description of health anxiety discourse on this subreddit in conjunction with the literature.

Future research may expand the baseline established here to explore variances and agreements. The present study can be used as a basis of comparison to contrast other online forums, time periods, mental health disorders, or changes in linguistic attributes for individual users over time. For example, tracking linguistic characteristics of users over time could yield some information related to possible efficacy of support forums if self-focus, negative emotion, and affect variables decrease. Further, opportunities exist to correlate LIWC categories with cognitive distortions that may vary in prevalence based on the pathology presented. Another possible avenue of exploration is investigating the impacts or effects of individuals who participate in forums and associative linguistic alterations in conjunction with outcome measures of health anxiety. Future research may evaluate the alterations in language use between a 2019 sample, as provided in this study, and language use during the COVID-19 pandemic. Also, exploring the impacts of those reading posts and comments may also be of benefit, such as investigating whether reading posts and comments increase empathy or understanding of the condition among counselors-in-training.

Limitations

The present study has several limitations. First, the linguistic attributes identified within the study populations represent monolingual English. Statistics on Reddit use for countries outside of the U.S. suggest minimal usage, and exact make-up of the nationality or English as a first language are not known within the sample population. As such, generalization to other nationalities, regions, or languages is not possible. Also, linguistic attributes may be generalized only to online communication for this group. The LIWC application cannot distinguish words in context and as such words used in various forms, such as sarcasm, warrant caution for generalizing results.

Further research is needed to compare results of in-person interpersonal communications for those with health anxiety. Generalizations may be limited as the comparison corpus was taken from general blog entries and not from other online forum formats. This study compared a 2019 study corpus to a 2012 reference corpus. As such, linguistic analysis may be impacted as new vocabulary may enter the lexicon over time (Leech, 2002). For example, proper noun usage may change due to current events. Further replication from different reference corpora before making definitive generalizations about the findings. Finally, generalization should be used with care as findings look at group-level linguistic attributes and may not account for an individual's specific linguistic attributes.

Conclusion

The current study sought to identify the linguistic attributes of those with health anxiety on the subreddit r/HealthAnxiety. The results describe a collection of communication that is authentic, high in self-focus, uses a significant amount of negative emotion words, and have higher levels of words associated with biological processes. The results support previous research regarding health anxiety. Further, the findings indicate that this subreddit may be a source for counseling professionals to reference to gain a deeper understanding of the experiences with health anxiety. Developing a deeper understanding of the experiences may provide greater insight into the internal worlds of those with health anxiety. Real-time posts or comments during acute episodes of health anxiety could provide an appreciation for the significant distress experienced by those with health anxiety. The current study can serve as a baseline for future research to compare linguistic attributes over time, with other groups, and in other online and offline contexts; it can also be used for user-specific studies.

References

Abramowitz, J. S., Deacon, B. J., & Valentiner, D. P. (2007). The short health anxiety inventory: Psychometric properties and construct validity in a non-clinical sample. *Cognitive Therapy and Research*, *31*(6), 871–883. https://doi.org/10.1007/s10608-006-9058-1

Al-Mosaiwi, M., & Johnstone, T. (2018). In an absolute state: Elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation. *Clinical Psychological Science*, *6*(4), 529–542. https://doi.org/10.1177/2167702617747074

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Author.

Asmundson, G. J. G., & Taylor, S. (2020). How health anxiety influences responses to viral outbreaks like COVID-19: What all decision-makers, health authorities, and health care professionals need to know. *Journal of Anxiety Disorders*, *71*, Article 102211. https://doi.org/10.1016/j.janxdis.2020.102211

Brezina, V. (2018). *Statistics in corpus linguistics: A practical guide*. Cambridge.

Coppersmith, G., Dredze, M., & Harman, C. (2014). Quantifying mental health signals in Twitter. *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 51–60. https://doi.org/10.3115/v1/W14-3207

Dean, B. (2021, February 25). *Reddit usage and growth statistics: How many people in use Reddit in 2021?* https://backlinko.com/reddit-user

De Choudhury, M., Counts, S., & Horvitz, E. (2013). Major life changes and behavioral markers in social media: Case of childbirth. *CSCW '13: Proceedings of the 2013 conference*

Computer Supported Cooperative Work, 1431–1442. https://doi.org/10.1145/2441776.2441937

De Choudhury, M., & De, S. (2014). Mental health discourse on Reddit: Self-disclosure, social support, and anonymity. *Proceedings of the International AAAI Conference on Web and Social Media*, *8*(1). https://ojs.aaai.org/index.php/ICWSM/article/view/14526

Everett, C. (2013). *Linguistic relativity: Evidence across languages and cognitive domains*. De Gruyter Mouton.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*(4)*,* 1149-1160. https://doi.org/10.3758/BRM.41.4.1149

Fischer-Homberger, E. (1972). Hypochondriasis of the eighteenth century—neurosis of the present century. *Bulletin of the History of Medicine*, *46*(4), 391–401.

Freud, S. (1960). *Psychopathology of everyday life*. E. Benn.

Furer, Walker, John R., & Stein, Murray B. (2007). *Treating health anxiety and fear of death : a practitioner's guide*. Springer. https://doi.org/10.1007/978-0-387-35145-2

Gaur, M., Kursuncu, U., Alambo, A., Sheth, A., Daniulaityte, R., Thirunarayan, K., & Pathak, J. (2018). “Let me tell you about your mental health!”: Contextualized classification of Reddit posts to DSM-5 for web-based intervention. *CIKM '18: Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 753–762. https://doi.org/10.1145/3269206.3271732

Greaves, M. M., & Dykeman, C. (2018). A corpus linguistic analysis of public Reddit blog posts on non-suicidal self-injury. *ArXiv*. https://doi.org/abs/1902.06689

Hadjistavropoulos, H. D., Janzen, J. A., Kehler, M. D., Leclerc, J. A., Sharpe, D., & Bourgault-Fagnou, M. D. (2012). Core cognitions related to health anxiety in self-reported medical and non-medical samples. *Journal of Behavioral Medicine*, *35*(2), 167–178. https://doi.org/10.1007/s10865-011-9339-3

Hardie, A. (2014, April 28). *Log ratio – an informal introduction.* http://cass.lancs.ac.uk/log-ratio-an-informal-introduction/

Holmes, E. A., O'Connor, R. C., Perry, V. H., Tracey, I., Wessely, S., Arseneault, L., Ballard, C., Christensen, H., Silver, R. C., Everall, I., Ford, T., John, A., Kabir, T., King, K., Madan, I., Michie, S., Przybylski, A. K., Shafran, R., Sweeney, A., … Bullmore, E. (2020). Multidisciplinary research priorities for the COVID-19 pandemic: A call for action for mental health science. *Lancet Psychiatry*, *7*(6), 547–560. https://doi.org/10.1016/S2215-0366(20)30168-1

Holtzman, N. S., Tackman, A. M., Carey, A. L., Brucks, M. S., Küfner, A. C. P., Deters, F. G., Back, M. D., Donnellan, M. B., Pennebaker, J. W., Sherman, R. A., & Mehl, M. R. (2019). Linguistic markers of grandiose narcissism: A LIWC analysis of 15 samples. *Journal of Language and Social Psychology*, *38*(5–6), 773–786. https://doi.org/10.1177/0261927X19871084

Ingram, R. E. (1990). Self-focused attention in clinical disorders: Review and a conceptual model. *Psychological Bulletin*, *107*(2), 156–176. https://doi.org/10.1037/0033- 2909.107.2.156

Leech, G. (2002). The importance of reference corpora. *Hizkuntza-corpusak. Oraina eta geroa*, *10*(24/25), 1-11. https://www.uzei.eus/wp-content/uploads/2017/06/06-Geoffrey-LEECH.pdf

Marcus, D. K., Hughes, K. T., & Arnau, R. C. (2008). Health anxiety, rumination, and negative affect: A mediational analysis. *Journal of Psychosomatic Research*, *64*(5), 495–501. https://doi.org/10.1016/j.jpsychores.2008.02.004

Mor, N., & Winquist, J. (2002). Self-focused attention and negative affect: A meta-analysis. *Psychological Bulletin*, *128*(4), 638–662. https://doi.org/10.1037//0033-2909.128.4.638

Newby, J. M., Smith, J., Uppal, S., Mason, E., Mahoney, A. E. J., & Andrews, G. (2018). Internet-based cognitive behavioral therapy versus psychoeducation control for illness anxiety disorder and somatic symptom disorder: A randomized controlled trial. *Journal of Consulting and Clinical Psychology*, *86*(1), 89–98. https://doi.org/10.1037/ccp0000248

O'Bryan, E. M., McLeish, A. C., & Johnson, A. L. (2017). The role of emotion reactivity in health anxiety. *Behavior Modification*, *41*(6), 829–845. https://doi.org/10.1177/0145445517719398

Park, A., & Conway, M. (2017). Longitudinal changes in psychological states in online health community members: Understanding the long-term effects of participating in an online depression community. *Journal of Medical Internet Research*, *19*(3), e71. https://doi.org/10.2196/jmir.6826

Pennebaker Conglomerates. (2021, October 14). *Interpreting LIWC output*. https://liwc.wpengine.com/interpreting-liwc-output/

Pennebaker, J. W. (2013). *The secret life of pronouns: What our words say about us*. Bloomsbury Press.

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC 2015*. University of Texas at Austin.

https://repositories.lib.utexas.edu/bitstream/handle/2152/31333/LIWC2015_LanguageManual.pdf?Sequence=3

Qiu, L., Lin, H., Ramsay, J., & Yang, F. (2012). You are what you tweet: Personality expression and perception on Twitter. *Journal of Research in Personality*, *46*(6), 710–718. https://doi.org/10.1016/j.jrp.2012.08.008

Rosnow, R. L., & Rosenthal, R. (2009). "Effect sizes for experimenting psychologists": Correction to Rosnow and Rosenthal (2003). *Canadian Journal of Experimental Psychology*, *63*(2), 123. https://doi.org/10.1037/a0015528

Rude, S., Gortner, E.-M., & Pennebaker, J. (2004). Language use of depressed and depression-vulnerable college students. *Cognition and Emotion*, *18*(8), 1121–1133. https://doi.org/10.1080/02699930441000030

R/Health Anxiety. (2021, June 20). R/HealthAnxiety. https://www.reddit.com/r/HealthAnxiety/

Saha, K., Torous, J., Caine, E. D., & De Choudhury, M. (2020). Psychosocial effects of the COVID-19 pandemic: Large-scale quasi-experimental study on social media. *Journal of Medical Internet Research*, *22*(11), e22600. https://doi.org/10.2196/22600

Salkovskis, P. M., & Warwick, H. M. C. (2001). Meaning, misinterpretations, and medicine: A cognitive-behavioral approach to understanding health anxiety and hypochondriasis. In V. Starcevic & D. R. Lipsitt (Eds.), *Hypochondriasis: Modern perspectives on an ancient malady* (pp. 202–222). Oxford University Press.

Schmidt, N. B., Joiner, T. E., Staab, J. P., & Williams, F. M. (2003). Health perceptions and anxiety sensitivity in patients with panic disorder. *Journal of Psychopathology and Behavioral Assessment*, *25*(3), 139–145. https://doi.org/10.1023/A:1023520605624

Semrush, 2021 (2022, February 7). Top 100: The most visited websites in the US. https://www.semrush.com/blog/most-visited-websites/

Shen Hanwen, J., & Rudzicz, F. (2017). Detecting anxiety on Reddit. *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology, Vancouver Canada*, 58–65.

Simmons, R. A., Chambless, D. L., & Gordon, P. C. (2008). How do hostile and emotionally overinvolved relatives view relationships?: What relatives' pronoun use tells us. *Family Process*, *47*(3), 405–419. https://doi.org/10.1111/j.1545-5300.2008.00261.x

Sonnenschein, A. R., Hofmann, S. G., Ziegelmayer, T., & Lutz, W. (2018). Linguistic analysis of patients with mood and anxiety disorders during cognitive behavioral therapy. *Cognitive Behaviour Therapy*, *47*(4), 315–327. https://doi.org/10.1080/16506073.2017.1419505

Starcevic, V., & Aboujaoude, E. (2015). Cyberchondria, cyberbullying, cybersuicide, cybersex: "New" psychopathologies for the 21st century? *World Psychiatry*, *14*(1), 97–100. https://doi.org/10.1002/wps.20195

Starcevic, V., & Berle, D. (2014). Cyberchondria: Towards a better understanding of excessive health-related Internet use. *Expert Review of Neurotherapeutics*, *13*(2), 205–213. https://doi.org/10.1586/ern.12.162

Tankovska, H. (2021a, May 3). *Distribution of Reddit app users in the United States as of March, 2021*. https://www.statista.com/statistics/1125159/reddit-us-app-users-age/

Tankovska, H. (2021b, April 21). *Regional distribution of desktop traffic to Reddit.com as of December 2020, by country.* https://www.statista.com/statistics/325144/reddit-global-active-user-distribution/

Tankovska, H. (2021c, May 3). *Percentage of U.S. adults who use Reddit as of February 2021, by gender.* https://www.statista.com/statistics/261765/share-of-us-internet-users-who-use-reddit-by-gender/

Tankovska, H. (2021d, June 11). *Percentage of U.S. adults who use Reddit as of February 2019, by ethnicity*. https://www.statista.com/statistics/261770/share-of-us-internet-users-who-use-reddit-by-ethnicity/

Taylor, S., & Asmundson, G. J. G. (2004). *Treating health anxiety: A cognitive-behavioral approach.* The Guilford Press.

Tyrer, P., Cooper, S., Tyrer, H., Wang, D., & Bassett, P. (2019). Increase in the prevalence of health anxiety in medical clinics: Possible cyberchondria. *International Journal of Social Psychiatry*, *65*(7–8), 566–569. https://doi.org/10.1177/0020764019866231

Wilson, A. (2013). Embracing Bayes factors for key item analysis in corpus linguistics. In M. Bieswanger, & A. Koll-Stobbe (Eds.), *New approaches to the study of linguistic variability* (pp. 3-11). (Language Competence and Language Awareness in Europe; Vol. 4). Peter Lang.

Walker, & Furer, P. (2008). Interoceptive Exposure in the Treatment of Health Anxiety and Hypochondriasis. *Journal of Cognitive Psychotherapy*, *22*(4), 366–378. https://doi.org/10.1891/0889-8391.22.4.366

Warwick, H. M. C., & Salkovskis, P. M. (1990). Hypochondriasis. *Behaviour Research and Therapy*, *28*(2), 105–117. https://doi.org/10.1016/0005-7967(90)90023-C

Weisser, M. (2016). *Practical corpus linguistics: An introduction to corpus-based language analysis*. John Wiley & Sons.

Yalom, I. D. (2013). *The gift of therapy: An open letter to a new generation of therapists and their patients*. Harper Perennial.

Chapter 3

Health Anxiety Posts on Reddit: A Keyness and Collocation Study

Christopher McBride

Kok-Mun Ng

Thom Field

Author Note

Christopher McBride, Counseling Academic Unit, Oregon State University; Kok-Mun Ng, Counseling Academic Unit, Oregon State University; Thomas Field, Counseling Academic Unit, Oregon State University.

The research contained in this manuscript was part of the first author's dissertation research project.

Correspondence concerning this article should be addressed to Christopher McBride, Email: mcbrichr@oregonstateu.edu

Abstract

Researchers are using computer assisted linguistic analysis software to provide information related to mental health conditions. By analyzing discourse from a variety of sources much can be learned about populations. Research in social media postings is providing information related to a variety of mental health conditions. Health anxiety is a symptom that overlaps a variety of mental health diagnoses. A study corpus was created from one-year of posts and comments from the subreddit r/HealthAnxiety from Reddit.com. The study corpus was compared to the blogs section from the corpora the Corpus of American English Corpus (The COCA, 2021). This study identified the top 100 keywords for the study corpus and the reference corpus. Subsequently, the most common collocates for the top five keywords and term "health anxiety" were identified for the study corpus. The results indicate that the uniqueness of the study corpus pertains to words regarding medications or diseases, medication and supplements, medical tests, online resources, symptom words, body part, and anxiety. A discussion of the results, clinical implications, counselor education implications, research, and limitations is included.

Keywords: health anxiety, keyness, keywords, collocates, AntConc, corpus linguistics

Health Anxiety Posts on Reddit: A Keyness and Collocation Study

Researchers' studying mental health disorders via social media interaction has increased over the past decade (Shen & Rudzicz, 2017). Utilizing a language analysis approach, such as corpus linguistics, allows researchers to investigate large data sets for linguistic patterns of interaction to gather insight into a user's psychology (Tausczik & Pennebaker, 2010). The ways in which individuals use words can provide information regarding beliefs, thinking patters, social relationships, personalities, and fears (Pennebaker et al., 2015). On social media websites, such as Reddit, discussion boards focus directly on mental health-related topics such as depression, suicide, and health anxiety. Despite the increase in research, keyword analysis and studies related to collocations have yet to be conducted on the subreddit related to health anxiety.

The prevalence of health anxiety is increasing (Tyrer et al., 2019). Those with health anxiety can experience symptoms for many years prior to diagnosis (Hedman et al., 2011). While health anxiety is often underdiagnosed, those with health anxiety can experience intense misery due to the condition and sometimes the result of the condition ends in suicide (Tyrer & Tyrer, 2018).

Health anxiety is fear of, preoccupation with, or negative interpretation of health-related symptoms, often out of proportion with the symptoms experienced (Salkovskis & Warwick, 2001). The *Diagnostic and Statistical Manual* (DSM-5) describes health anxiety symptoms within the somatoform disorders or illness anxiety disorder (American Psychiatric Association, 2013). Health anxiety may be experienced on a spectrum from mild to severe symptoms. In more severe cases, individuals may struggle with persistent ideation of a fatal disease without appropriate physiological symptoms to substantiate this preoccupation (Warwick & Salkovskis, 1990). While specific disorders address health anxiety specifically, its symptoms are often

components of other anxiety disorders such as generalized anxiety disorder, obsessive-compulsive disorder, and panic disorder (Abramowitz et al., 2007). Research related to individuals' linguistic qualities with health anxiety may increase a practitioner's ability to recognize the experience and common language of those who meet diagnostic criteria.

The cognitive behavior model of health anxiety provides a comprehensive explanation for health anxiety and is well researched (Taylor & Asmundson, 2004; Warwick & Salkovskis, 1990). Factors for development of health anxiety are biological predisposition to anxiety, stressful life events, or exposure or experiencing difficult illness or death-related experience (Walker & Furer, 2008). The cognitive behavior model of health anxiety describes a cycle consisting of triggers, emotions, and behaviors. Health anxiety may be elicited by the individual experiencing either an internal or external trigger (Furer et al., 2007; Walker & Furer, 2008). Following trigger events, those with health anxiety experience an illness thought accompanied with anxiety or fear, increased bodily sensations, and subsequent overestimation of the severity of bodily sensations. Next, those with health anxiety engaged in coping behaviors which may be increased checking, reassurance seeking, avoidance, medical appointments, and so forth. While coping strategies may temporarily diminish anxiety, coping behaviors will often maintain the experience of health anxiety as the anxiety cycle waxes and wanes (Warwick & Salkovskis, 1990).

Interestingly, the emergence and common use of the Internet is impacting the rates of health anxiety. Tyrer et al. (2019) reported an increase in health anxiety from 14.9% in 2006-2008 to 19.9% in 2008-2010; self-diagnosis from web-based interactions contributing to the increase. As health-related interactions increase with online tools, the emergence of cyberchondria is a condition of note. Cyberchondria is defined as repeated or excessive searching

of health-related information via online resources (Starcevic & Aboujaoude, 2015; Starcevic & Berle, 2014). More extended periods of internet use focused on health-related information correlates with increased levels of anxiety. Paradoxically, individuals seek to reduce anxiety via health-related and symptom-related searches. However, users often discover new information, which may increase rumination and anxiety. Understanding how those with health anxiety use the Web and interact with others is an essential topic for mental health researchers to investigate.

Individuals with health anxiety may find value in engaging with an online community. Reeves et al. (2014) found that individuals with chronic illnesses benefit from social involvement with online group interactions. The researchers noted that engagement with internet-based technology (e.g., community groups) might buttress self-management and mental and physical wellbeing. Further, individuals might grow personal networks as their health needs evolve. In addition, support garnered from social network interactions may substitute for formal care, which reduces healthcare costs. Despite these findings, little is currently known about users' experiences with health anxiety on social media networks such as Reddit. Researchers can gain further understanding by utilizing a corpus linguistics approach.

Reddit is a social media platform with 430 million active users (Dean, 2021). As of December 2021, it was listed as the seventh most visited website in the United States (Semrush, 2021). Users are called "redditors." Redditors may join communities, post and comment, upvote or downvote comments, create topic specific subreddits, and interact with the website in various other ways. The primary language on Reddit is English without a site-based mechanism to change the language. However, specific subreddits may be created for non-English speakers.

The userbase of Reddit is varied. As of February 2021, 18-29-year-olds comprised 36% of the redditors, 30-49-year-olds comprised 22%, 50-64-year-olds comprised 10%, and 65 and

older comprised 3% (Tankovska, 2021a). The United States comprises the largest portion of redditors at 49.32%, then the United Kingdom at 7.85%, Canada at 7.76%, Australia at 4.34%, and Germany at 3.11% (Tankovska, 2021b). Within the United States, 12% of females and 23% of males were reported using Reddit; nonbinary genders were not reported (Tankovska, 2021c). Tankovska (2021d) reported that of U.S. residents, 4% of Black Americans, 14% of Hispanics, and 12% of White Americans use Reddit.com; other racial and ethnic groups were not reported. Demographic data for specific subreddit usage is not known at this time.

A corpus represents a text or a set of texts, which can take either oral or written form, identified for linguistic analysis (Weisser, 2016). Researchers have a great deal of flexibility in the construction of a corpus. Corpus linguistics researchers focused on contemporary speech may conduct synchronic design. Conversely, researchers interested in historical or comparative studies may adopt a diachronic design. Researchers could construct a static corpus in that the corpus is fixed and not open to change, or they can adopt a dynamic approach in which the corpus changes over time to reflect the evolving nature of language (Weisser, 2016).

A reference corpus is a collection of written or spoken words by which a study corpus is compared. A reference corpus should be comprised of source material that provides general information about the language (Leech, 2002). Leech asserts that the reference corpus should be viewed as a bench mark by which to compare other uses of language. A reference corpus should represent multiple sources to ensure diversity. The Corpus of Contemporary American English is one of the most widely used corpora for the English language (COCA; The COCA, 2001). The COCA blogs section includes blogs postings as identified by Google.com from academic, argument, fiction, info, instruction, legal, news, personal, promotion, and review web pages.

Considering the aforementioned, the COCA blogs section meets the criteria forwarded by Leech (2002) for use in corpus linguistic analysis.

Little is known about the experiences of those with health anxiety as it presents via online forums such as Reddit. Reddit is one of the largest, free, online forums where it has been observed that users express their views about topics from a wide range of subject matter. It has been noted that due to the significant character limits and anonymity, users are more expressive and open about mental health experiences (De Choudhry & De, 2014). Shen and Rudzicz (2017) investigated anxiety across several Reddit groups or subreddits. They identified that anxiety, negative, dictionary words, and affect were among the most common types of terms used by users. Further, the researchers identified unigrams, bigrams, and trigrams common across groups of anxiety subreddits compared to a group of selected control subreddits. However, results specific to the health-anxiety subreddit were not identified within their study.

Given the aforenoted gaps and needs, this study sought to identify keywords and collocations existing within the subreddit r/HealthAnxiety. The research questions were:

1. What words distinguish online posts about health anxiety from online posts in general?
2. What words distinguish general online posts from online posts about health anxiety?
3. What are the most common collocations of the strongest keywords of online posts about health anxiety?
4. What are the most common collocations of the term “health anxiety” in online posts about health anxiety?

There is a compelling argument for researchers in counseling and related fields to use corpus linguistic methods. First, Pennebaker (2012) suggested that a revolution is underway with

linguistic analysis that will have a profound impact on the social sciences. Pennebaker observed that words are like fingerprints which provide information related to an individual's identity. Specific to counseling, LaGue et al. (2019) assert that corpus linguistic tools can provide insight into unique worldviews, provide objective data regarding effective interventions, and may contribute to skill development for counselors-in-training (p. 13). Greaves and Dykeman (2018) studied the linguistic attributes for individuals with nonsuicidal self-injury; they stated that understanding and analyzing language may help counselors provide information related to their inner experience. The researchers asserted that such information would assist the mental health field in understanding individuals regarding their pathology better. Regarding these observations mentioned above, language analysis and corpus linguistic methodology are appropriate for research within the counseling and mental health fields.

Method

Design

The present research study used a synchronic corpus linguistic design to observe the characteristics of health anxiety posts on Reddit (Weisser, 2016). The Corpus of Contemporary American English was used as the reference corpus (COCA; The COCA corpus, 2021). The application AntConc 3.5.9 was used to identify the top keywords and collocations within the corpus (Anthony, 2020). Keywords and collocations indicate unique attributes of the corpus of interest and provide insight into the corpus of references' "aboutness" (Egbert & Biber, 2019).

Power Analysis

An a priori power analysis was completed for the research questions using G*Power (Faul et al., 2009). The tests used are a derivation of the $\chi 2$ test. The appropriate effect size is Cohen's *w*; we assumed a medium effect of 0.30 (Rosnow & Rosenthal, 2003). The input

parameters were: (a) test-family- $\chi 2$ test; (b) statistical test- goodness-of-fit tests: contingency tables; (c) type of power analysis – a priori: compute required sample size- given, power, and effect size; (d) $w = 0.3$; (e) power (1- β error probability) = 0.90; (f) $\alpha = .05$; and (g) degrees of freedom (*Df*) = 1. The G*Power output indicated a sample size of 117 for an actual power of 0.90.

Study Corpus

Register, Scope, and Sources. The register is internet-based informal discourse. The subregister is health anxiety communication. The scope is comments and replies on a public forum dedicated to health anxiety. The source is the subreddit r/Healthanxiety for the calendar year 2019. The year was selected as the COVID-19 impacted most of 2020, potentially influencing normative communication on a health anxiety board. A full calendar year was selected to control for seasonal impacts on communication.

Preprocessing. Raw data from comments and replies were extracted from the r/HealthAnxiety subreddit. Data were output in text format. Usernames were not gathered to protect users' privacy. Direct quotations with personal identifying information were not used in the results or discussion sections of this study. Colloquial words and terms were converted to standard English (e.g., "lol" to laugh out loud, "btw" to by the way, "HA" to health anxiety, "IMHO" to in my humble opinion). Punctuation was added as necessary and wherever possible (e.g., "ive" to "I've," "im" to "I'm," and so forth). Stop words were removed to reduce noise within the corpus.. The text files were merged as necessary to prepare for processing by the AntConc application.

Reference Corpus

Register, Scope, and Sources. The reference corpus selected is the COCA blog portion (The COCA Corpus, 2021). The register was online informal discourse. The scope is blog posts identified by Google.com covering a variety of topics including academic, argument, fiction, info, instruction, legal, news, personal, promotion, and review web pages for the year of 2012 (The COCA Corpus, 2021). The source is the blog portion of the COCA.

Preprocessing. The preprocessing strategy was undertaken to the reference corpus as that of the study corpus. Colloquial words and terms were converted to standard English (e.g., "lol" to laugh out loud, "btw" to by the way, "IMHO" to in my humble opinion). Stop words were removed to reduce noise within the corpus. COCA text files were merged into a single file for preprocessing. After preprocessing, due to the file size, the file was divided into nine separate files for use in AntConc when using it as a comparison for the study corpus.

Measures

Keyness. This study sought to identify keywords and collocates within the subreddit r/HealthAnxiety. A keyword list was used to identify the *keyness* of a text. Keyness is a common corpus linguist approach to determine the "aboutness" of a text (Egbert & Biber, 2019). Keywords are words that appear more frequently within the study corpus than the reference corpus. These keywords represent the uniqueness of the corpus.

Node Word. A node word is a central word, phrase, or grammatical structure that is central to how co-occurrence patterns are evaluated (Brezina et al., 2018; McEnery & Hardie, 2011). As such, node words are vital to the identification of collocations, and they represent the central search terms used. For this study, node words were the top five keywords as well as the term "health anxiety."

Stop words. Words that do not provide information or substance specific to the subject of interest are considered stop words (Geisler & Dykeman, 2021). For this study, removing stop words decreased the noise within the corpus. Examples of stop words are "the," "is," "and," and "as." The National Language Toolkit (NLTK) list was used to identify which stop words to remove from the study and reference corpus (Loper & Bird, 2002).

Collocation. Collocations provide additional information related to an interpretation of keywords within a corpus. Firth (1968), a significant figure in linguistics, wrote regarding collocations, "You shall know a word by the company it keeps" (p. 179). Brezina (2018) asserted that knowledge related to essential words within a corpus may be enhanced by understanding words that appear in close proximity.

Apparatus

AntConc 3.5.9 is a text analysis software that can determine lexical patterns from a corpus or corpora (Anthony, 2020). AntConc is used to identify concordance, clusters, n-grams, collocation, word lists, and keyword lists. Users may select a reference corpus to compare a corpus of interest within the user interface.

Data Analysis

The AntConc application was used to identify the top 100 keywords for the corpus of interest compared to the reference corpus. The following were reported: rank, frequency, keyness, effect, and keyword. Within keyword identification, two significant metrics are typically required: significance and effect (Gabrielatos, 2018). This study used the log-likelihood test to determine statistical significance. Traditionally, in the social sciences, *p* values of less than .05 have been deemed sufficient. However, because of the multiple tests that are conducted as part of keyness studies, the keyword statistic is often assumed at far more conservative *p*

values, with $p < .01$. Hardie's (2014) log-ratio test was used to test for effect size. Bonferroni correction was used to control for family-wise error rate. Parameters for keyness values were set as "log-likelihood 4-term," keyword statistic threshold "$p < .01$ + Bonferroni," keyword effect size measure "Hardie's Log Ratio," key effect size threshold "Top 100."

For collocations, the top five keywords and the term "health anxiety" were identified as node words. AntConc analyzed the collocation stems for frequently occurring terms within the corpus. The parameters were set as search terms = "words," from "5L" to "5R," minimum collocation frequency = "3," sort by "frequency." The paraments were set as search terms = "words," from "5L" to "5R," minimum collocation frequency = "3," sort by "frequency." The top five results for each collocation are reported. The mutual information (MI) statistic was to determine effect size. MI scores of three or greater are considered of linguistic importance (Hunston, 2002).

Results

The study corpus was constructed from comments and replies from r/HealthAnxiety for the year 2019. For corpus studies a unique word form is referred to as a "type" and a single occurrence of a word is referred to as a "token" (Brezina, 2018). Total comments and replies were 80,129. For the reference corpus, total entries were 98,796. After preprocessing, a total 2,807,341 tokens were in the study corpus, and a total of 56,483,561 tokens were in the reference corpus.

A keywords analysis within the AntConc 3.5.9 application was comparing the study corpus to the reference corpus. Log-likelihood (G^2) was used to determine significance. The top 100 keywords in the study corpus proved to be statistically significant (G^2 = 9127.9 to 94.26, df = 1, $p < .01$). Hardie's (2014) Log-ratio showed a substantial effect size (LR = 13.381 to 9.1715).

Every point above zero equals a doubling in size in comparing the frequency of a keyword in comparing the two corpora (Hardie, 2014). Interpreted, the log-ratio scores indicates that keywords range approximately 14,304.304 to 687.616 times more common in the study corpus than in the reference corpus.

For the study corpus, the results can be viewed in Table 3.1.

Table 3.1

Keyword for Study Corpus Results (RQ 1)

Rank	Keyword	Frequency	G^2	Log-ratio
1	palpitations	1610	9127.9**	9.5764
2	amyotrophic	1300	7562.25**	11.5158
3	ekg	852	4816.3**	9.4468
4	lightheaded	344	1958.63**	9.7904
5	hypochondria	333	1910.22**	10.229
6	dvt	332	1896.23**	9.9616
7	ekgs	296	1743.76**	13.381
8	keto	268	1507.25**	9.2376
9	holter	245	1380.1**	9.3008
10	tonsil	226	1276.33**	9.4067
11	temporomandibular	215	1253.94**	11.9197
12	pvcs	213	1242.18**	11.9062
13	emg	200	1155.99**	10.8154
14	healthanxiety	157	924.89**	12.4661
15	echocardiogram	151	877.61**	11.4099
16	vaping	149	848.6**	9.8057
17	healthycliq	143	842.42**	12.3314
18	costochondritis	140	824.74**	12.3008
19	vape	140	803.94**	10.3008
20	derealization	136	801.18**	12.259
21	palpitation	130	745.32**	10.1939
22	nembutal	126	742.27**	12.1488
23	propranolol	123	713.07**	11.114
24	ecgs	108	636.23**	11.9264
25	utm	106	624.45**	11.8994

Rank	Keyword	Frequency	G^2	Log-ratio
26	globus	106	604.75**	9.8994
27	buspar	100	589.1**	11.8154
28	lymphnodes	100	589.1**	11.8154
29	sertraline	99	583.21**	11.8009
30	askdocs	93	547.86**	11.7107
31	lipoma	89	524.3**	11.6473
32	palps	91	516.98**	9.6793
33	omeprazole	87	493.6**	9.6145
34	hiatal	74	435.93**	11.381
35	fowleri	70	412.37**	11.3008
36	dimer	69	396.1**	10.2801
37	petechiae	65	372.66**	10.1939
38	juul	62	355.08**	10.1257
39	flonase	58	331.65**	10.0295
40	naegleria	55	324.01**	10.9529
41	emetophobia	53	312.22**	10.8994
42	escitalopram	52	306.33**	10.872
43	gastroscopy	53	302.37**	9.8994
44	producthunt	48	282.77**	10.7565
45	lymphnode	46	270.99**	10.6951
46	diverticulitis	47	267.26**	9.7261
47	fasciculation	44	259.2**	10.631
48	lipomas	44	259.2**	10.631
49	miralax	43	243.88**	9.5978
50	fasciculations	40	235.64**	10.4935
51	medicatedpharmacyonline	40	235.64**	10.4935
52	buspirone	39	229.75**	10.4569
53	palpations	39	229.75**	10.4569
54	dpdr	33	194.4**	10.2159
55	tanax	33	194.4**	10.2159
56	tetracaine	33	194.4**	10.2159
57	bulbar	31	182.62**	10.1257
58	metoprolol	31	182.62**	10.1257
59	prostatitis	29	170.84**	10.0295
60	unruptured	29	170.84**	10.0295
61	catastrophizing	27	159.06**	9.9264
62	alkatone	26	153.17**	9.872
63	hypnic	26	153.17**	9.872

Rank	Keyword	Frequency	G^2	Log-ratio
64	bradycardia	25	147.27**	9.8154
65	hemmorhoids	25	147.27**	9.8154
66	anxietycentre	24	141.38**	9.7565
67	sneakpeekbot	24	141.38**	9.7565
68	bionatrol	23	135.49**	9.6951
69	anxietys	22	129.6**	9.631
70	ashwagandha	22	129.6**	9.631
71	bodystart	22	129.6**	9.631
72	epididymitis	22	129.6**	9.631
73	heartfailure	22	129.6**	9.631
74	roemheld	22	129.6**	9.631
75	vaped	22	129.6**	9.631
76	euthasol	21	123.71**	9.5638
77	luhukpyuyonn	21	123.71**	9.5638
78	myehgcy	21	123.71**	9.5638
79	nutrafitz	21	123.71**	9.5638
80	setraline	21	123.71**	9.5638
81	zlzxztrr	21	123.71**	9.5638
82	exitunit	20	117.82**	9.4935
83	ppxvku	20	117.82**	9.4935
84	bimyyl	19	111.93**	9.4195
85	chrons	19	111.93**	9.4195
86	dvts	19	111.93**	9.4195
87	gluco	19	111.93**	9.4195
88	hyperfocusing	19	111.93**	9.4195
89	lymes	19	111.93**	9.4195
90	emgs	18	106.04**	9.3415
91	hhwqi	18	106.04**	9.3415
92	nexplanon	18	106.04**	9.3415
93	oraquick	18	106.04**	9.3415
94	purabella	18	106.04**	9.3415
95	manplus	17	100.15**	9.259
96	raynauds	17	100.15**	9.259
97	supplementgo	17	100.15**	9.259
98	tonsilitis	17	100.15**	9.259
99	vixea	17	100.15**	9.259
100	cardiophobia	16	94.26**	9.1715

Note: The G^2 critical value for $^{**}p < .01$ is 6.63. $df = 1$.

A keywords analysis within the AntConc 3.5.9 application was comparing the reference corpus to the study corpus. The top 100 keywords in the reference corpus proved to be statistically significant (G^2 = 6901.08 to 295.55, df = 1, p < .01). Hardie's (2014) log-ratio ranged showed differences in frequency in keywords for the reference corpus ((LR = 12.7908 to 8.2464). Interpreted, the log-ratio scores indicates that keywords range approximately 7355.11 to 378.88 times more common in the reference corpus than in the study corpus.

the results can be viewed in Table 3.2.

Table 3.2

Keyword for Reference Corpus Results (RQ 2)

Rank	Keyword	Frequency	G^2	Log-ratio
1	obama	63845	6901.08**	12.7908
2	toolong	41259	4458.7**	12.1609
3	president	38967	4179.18**	10.0785
4	romney	33143	3581.33**	11.8449
5	political	27388	2928.97**	9.5698
6	election	23096	2495.43**	11.3239
7	gt	22084	2356.6**	9.2592
8	federal	17312	1870.38**	10.908
9	economic	17327	1843.51**	8.9092
10	economy	17095	1818.5**	8.8898
11	republican	16335	1764.81**	10.8242
12	israel	16464	1763.22**	9.8355
13	campaign	16263	1728.79**	8.8178
14	republicans	13580	1467.12**	10.5577
15	global	13688	1451.24**	8.5691
16	taxes	12121	1282.42**	8.3937
17	voters	11293	1220.01**	10.2916
18	congress	11122	1201.54**	10.2696
19	democrats	11106	1199.81**	10.2676

Rank	Keyword	Frequency	G^2	Log-ratio
20	voting	10017	1067.62**	9.1187
21	director	9290	1003.6**	10.01
22	teams	9386	999.57**	9.0248
23	democratic	9040	976.59**	9.9706
24	leaders	8706	926.25**	8.9163
25	gop	8086	873.52**	9.8097
26	mitt	8030	867.47**	9.7997
27	policies	7838	832.69**	8.7648
28	businesses	7421	801.68**	9.6859
29	presidential	7275	785.91**	9.6572
30	george	6993	741.62**	8.6002
31	senate	6777	732.1**	9.5549
32	candidates	6763	716.84**	8.552
33	james	6703	710.38**	8.5391
34	iraq	6686	708.55**	8.5354
35	muslim	6478	699.8**	9.4898
36	flickr	6230	673.01**	9.4335
37	san	6352	672.57**	8.4615
38	jews	6138	663.07**	9.4121
39	clinton	5964	644.27**	9.3706
40	smith	5945	642.22**	9.366
41	committee	5883	635.52**	9.3508
42	politicians	5881	635.31**	9.3504
43	microsoft	5439	587.55**	9.2376
44	writers	5477	578.33**	8.2477
45	elections	5341	576.97**	9.2114
46	democracy	5324	575.13**	9.2068
47	reform	5289	571.35**	9.1973
48	christians	5225	564.44**	9.1797
49	israeli	5164	557.85**	9.1628
50	conservatives	5092	550.07**	9.1425
51	barack	4933	532.89**	9.0968
52	muslims	4906	529.97**	9.0888
53	voter	4732	511.18**	9.0367
54	editor	4717	509.56**	9.0322
55	democrat	4654	502.75**	9.0128
56	historical	4650	502.32**	9.0115
57	talent	4585	495.3**	8.9912

Rank	Keyword	Frequency	G^2	Log-ratio
58	fiscal	4528	489.14**	8.9732
59	amendment	4292	463.64**	8.8959
60	liberals	4089	441.71**	8.826
61	liberty	4039	436.31**	8.8083
62	ceo	4033	435.66**	8.8061
63	mccain	4014	433.61**	8.7993
64	unions	4009	433.07**	8.7975
65	corporations	3970	428.86**	8.7834
66	revolution	3933	424.86**	8.7699
67	gaza	3930	424.54**	8.7688
68	legislation	3850	415.89**	8.7392
69	investors	3690	398.61**	8.6779
70	estate	3477	375.6**	8.5921
71	innovation	3422	369.66**	8.5691
72	dan	3336	360.37**	8.5324
73	ballot	3315	358.1**	8.5233
74	racism	3273	353.56**	8.5049
75	blogging	3209	346.65**	8.4764
76	lee	3187	344.27**	8.4665
77	reagan	3153	340.6**	8.451
78	mortgage	3128	337.9**	8.4395
79	hamas	3121	337.14**	8.4363
80	palestinian	3087	333.47**	8.4205
81	minister	3078	332.5**	8.4163
82	developers	3045	328.93**	8.4007
83	andrew	3009	325.04**	8.3836
84	draft	3005	324.61**	8.3817
85	constitutional	2991	323.1**	8.3749
86	cia	2988	322.77**	8.3735
87	williams	2981	322.02**	8.3701
88	palestinians	2926	316.08**	8.3432
89	immigration	2915	314.89**	8.3378
90	benghazi	2909	314.24**	8.3348
91	crimes	2873	310.35**	8.3169
92	hollywood	2866	309.59**	8.3133
93	empire	2830	305.71**	8.2951
94	socialist	2826	305.27**	8.2931
95	enterprise	2817	304.3**	8.2885

Rank	Keyword	Frequency	G^2	Log-ratio
96	jeff	2808	303.33**	8.2838
97	capitalism	2804	302.9**	8.2818
98	senator	2793	301.71**	8.2761
99	politically	2787	301.06**	8.273
100	obamacare	2736	295.55**	8.2464

Note: The G^2 critical value for $^{**}p < .01$ is 6.63. $df = 1$.

For research questions three and four, the analysis was completed in AntConc 3.5.9 (Anthony, 2020). The node words were derived from the top five keywords in the study corpus. The node words were "palpitations," "amyotrophic," "ekg," "lightheaded," "hypochondria," and the term "health anxiety." Mutual information (MI) was used to identify collocates that are of linguistic interest (Hunston, 2002). MI scores ranged from 11.06028 to 2.31687. Twenty-six out of the 30 collocates exceeded the minimum threshold for linguistic importance. The most frequent collocates, frequency, frequency left, frequency right, and their corresponding MI scores are visible in Table 3.3.

Table 3.3

Collocation Results (RQ 3-4)

Node Word	Rank	Frequency	Freq. L	Freq. R	MI	Collocate
Palpitations	1	968	782	186	7.02455**	heart
	2	344	166	178	3.99511**	anxiety
	3	267	183	84	4.23095**	get
	4	197	72	125	3.38433**	like
	5	179	72	107	3.05074**	significant
Amyotrophic	1	1560	182	1378	10.37317**	sclerosis
	2	1380	54	1326	11.06028**	lateral
	3	172	105	67	7.41339**	multiple
	4	137	103	34	2.97543	anxiety
	5	110	50	60	2.65683	significant
ekg	1	232	118	114	5.8818**	heart
	2	210	68	142	6.33306**	blood

Node Word	Rank	Frequency	Freq. L	Freq. R	MI	Collocate
	3	148	37	111	5.91468**	normal
	4	135	53	82	6.3557**	chest
	5	124	34	90	6.69945**	done
Lightheaded	1	102	70	32	5.33708**	feel
	2	93	54	39	6.3717**	feeling
	3	87	54	33	9.30832**	dizzy
	4	76	56	20	4.64476**	get
	5	72	26	46	4.15878**	like
Hypochondria	1	101	58	43	4.50053**	anxiety
	2	51	24	27	3.51282**	significant
	3	49	24	25	4.82009**	health
	4	33	21	12	4.24377**	symptoms
	5	31	18	13	5.05478**	people
Health Anxiety	1	10802	775	10027	5.8804**	anxiety
	2	1387	664	723	2.91721	significant
	3	799	447	352	2.31687	like
	4	769	505	264	3.0819**	know
	5	670	357	313	3.10984**	really

Note: Freq. L is frequency left. Freq. R is frequency right. **MI is Mutual information, MI cut-off of 3 or more is considered worthy of linguistic importance (Hunston, 2002).

Per the results of the top 100 keyword categories, a pattern emerged. The word categories of anxiety word, symptom, body part, medication condition or disease, medical test, medication or supplement, other, online or other resource, personal habit, spam/automated message, and symptom word were identified.

To verify these categories, one doctoral candidate and two doctoral-level clinicians verified the results by individually categorizing each word. The interraters agreed that the categories provided adequate coverage for the top 100 keywords with only 1% of words falling under the category of "other." The interraters had an 83.76% agreement during the initial word categorization. They met to discuss the word categorization and reevaluate individual

categorizations, after which the word categorization agreement was 99.66%. This can be viewed in Table 3.4.

Table 3.4

Keyword Category Results

Word Category	Frequency	Percentage of Category
Medical condition or disease	30	29%
Medication or supplement	22	22%
Medical test	10	10%
Online or other resource	9	9%
Symptom word	7	7%
Spam/Automated word	7	7%
Personal habit	5	5%
Body part	5	5%
Anxiety word	4	4%
Other	1	1%

Note: Categories derived from words in the top 100 keyword list for study corpus.

Discussion

Yalom (2013) emphasized counselors entering the client's internal world. According to Pennebaker et al. (2015), words reveal important information related to their beliefs, thinking patterns, social relationships, personality, and fears. Corpus linguistic tools and techniques enable counseling professionals to view the dialog of those with varying pathologies at a group level. Doing so can increase the understanding of individuals who experience anxiety,

depression, and other mental health conditions. The corpus linguistic approach can provide counselors with another perspective to enter the world of individuals who suffer health anxiety.

The purpose of the current study was to examine the most frequently used words and associative words. The current study investigated what was being communicated that is unique to the subreddit r/HealthAnxiety. As a result, the top 100 keywords were identified as well as the top collocates for the most frequent keywords and the term "health anxiety" compared to a reference corpus. Per the results of the top 100 keywords for the study corpus, a pattern emerged regarding the types of words used. In the analysis, certain word categories emerged. Redditors used words pertaining to anxiety words or symptoms, body parts, medical conditions or diseases, medical tests, medication and supplements, online resources, personal habits, and symptoms. Also included in the keywords were words associated with automated or spam listings.

To evaluate the keywords and word categories, the CBT model for health anxiety provides a possible insight into the keywords within the corpus. The CBT model describes a cycle that originates from internal or external triggers leading to anxiety and illness thoughts (Furer et al., 2007). Before analysis of possible intersections between the CBT cycle and keywords and collocates, it is essential to note keywords are not identified in context via a collocation analysis, so it is beyond the scope of this study to determine whether words within the results originate more frequently from original posts that often describe an anxiety event, or reassurance comments. However, collocation analysis provides an additional layer of context to determine how the top five keywords and the term "health anxiety" are being applied within the corpus.

Keywords associated with body parts and symptoms may be indicative of initial internal triggers described within the CBT cycle. Body parts and symptoms words may play a significant

role as those with health anxiety often misinterpret internal and external triggers as being a sign of a more significant medical condition (Furer et al., 2007). Internal triggers lead to illness thoughts and anxiety or fear. For example, the most prevalent keyword "palpitation," as in "heart palpitation," could be a manifestation of the internal trigger experienced to begin the health anxiety cycle. Collocation analysis provides further context: at least a portion of "palpitation" word usage appears to originate from an experiential description. The association with anxiety is manifested in the collocate "anxiety" and "significant" which occur close to the node word. The aforementioned seems to align with the CBT model. "Palpitation" may either be an internal trigger or a physiological symptom experienced after anxiety or fear.

The symptom word of "lightheadedness" appears as the fourth most common keyword for the study corpus. "Lightheadedness" may be lesser known symptom associated with an anxiety response. Abramowitz & Deacon (2004) note that this can be a symptom associated with health anxiety. The researchers assert that "lightheadedness" may occur as a result of activation of an autonomic nervous system response, specifically associated with hyperventilation. Tyrer & Tyrer (2018) assert that health anxiety is often underdiagnosed. Clinicians may find the inclusion of lesser known symptoms, such as "lightheadedness," to be useful in identifying potential indicators of health anxiety.

Anxiety keywords may be representations of the distress experienced after the internal or external trigger or the exaggeration of symptoms as described in the CBT model. For example, the abbreviation "dpdr," which stands for depersonalization and derealization, may be indicative of the levels of distress caused by the experienced anxiety leading to posttraumatic type of symptoms. This aligns with the description of significant distress experienced for those with illness anxiety and somatoform disorders in the *DSM-5* (American Psychiatric Association,

2013; Tyrer & Tyrer, 2018). However, those symptoms are not listed as criteria for illness anxiety disorder or somatoform disorder. The presence of this keyword may be indicative of more extreme levels of health anxiety—that the distress of the anxiety causes symptoms aligned with a dissociative response. Perhaps the symptom may be indicative of comorbidity between health anxiety and trauma response or dissociative pathology. Traumatic experiences leading to onset of health anxiety corresponds with Walker & Furer's (2008) description of onset possibly being a result of a previous serious illness or death-related experience. Illness anxiety disorder is listed as a differential diagnosis is the *DSM-5* for depersonalization/derealization disorder. It is unclear whether the keyword "dpdr" describes a comorbidity with depersonalization/derealization disorder, or whether trauma-type symptoms are being described.

Several anxiety tokens may be conceptualized within the reassurance, checking part of the CBT model. Keywords such as "hyperfocusing" and "catastrophizing" are other anxiety terms that may indicate self-awareness of the user's anxiety disorder and a means to rationalize their current experience. Or, perhaps, those terms emanate from replies to posts by other redditors as an attempt to explain and reassure the original poster. "Hypochondriasis" is the fifth highest keyword, and "anxiety" is the most frequent collocate, followed by "significant." This suggests that the keyword "hypochondriasis" is used, at least some of the time, in context of a personal descriptive portion of the text. Also, investigating the term "health anxiety," collocates include "anxiety," "know," and "really." "Know" as a collocate could be additional information regarding insight the redditor experiences of their condition and "really" may be indicative of the intensity experienced.

In terms of overestimation of medical conditions described in the anxiety cycle, keywords within the top 100 list may lead to further evidence of this phenomena. Keywords such

as “fowleri” correspond with naegleria fowleri, which is commonly known as a brain eating ameba. Interestingly, according to the Center for Disease Control and Prevention (2017), only 151 individuals have experienced this condition in the United States since 1960. However, this word appears 70 times within a one-year span on the subreddit. Additionally, the keyword “amyotrophic,” which is a part of amyotrophic lateral sclerosis, commonly known as ALS or Lou Gehrig’s disease, is a common term describing a progressive and fatal medical condition. The description of this condition is consistent with the collocates “sclerosis” and “lateral.”

The above terms may be reflective of the interpretation of internal sensations of signs of serious illness. This interpretation of serious illness may be linked to catastrophizing. Marcus et al. (2018) found a correlation between catastrophizing, rumination, health anxiety, and negative affect. It is possible that words represent the thoughts and conclusions developed in this part of the cycle and are later communicated in the reassurance-seeking part of the cycle as redditors describe their health anxiety experiences within the online forum. In contrast to words regarding fatal medical conditions, other medical terms suggest more benign explanations.

At a global level, the act of posting on r/HealthAnxiety may be conceptualized within the reassurance seeking portion of the anxiety cycle. Further, keywords such as “costochondritis” offer more benign explanations for certain symptoms such as chest pain. It is possible that keywords with benign explanation are attempts for either the poster to self-reassure or for commentors to offer reassurance. Keywords related to reassurance may also be located within the categories of online resources or other resources words. Several terms exist within the board that are URL terms to online resources for self-help videos, or other subreddits such as “askdoc.” These keywords seem to reflect information and resource sharing consistent with the stated purpose of the subreddit to offer “A place for people with Health Anxiety / Illness Anxiety /

Hypochondria to come together and start taking control of their disorder" (r/Health Anxiety, 2021).

Medications and supplement keywords may factor into the reassurance part of the CBT model. Redditors may describe the medications and supplements utilized to combat anxiety or increase health. Keywords such as "buspar" and "buspirone" are anti-anxiety medications. Interestingly, a fair amount of "spam" posts exists on the subreddit, where companies advertise various products related to health. Of concern are two products within the top 100 keyword list: "nembutal" and "tanax." These two products are used for euthanasia in animals and commonly used for suicide attempts.

The current research represents, to our knowledge, the first investigation into keywords and collocates from an online forum regarding health anxiety. These keywords provide additional insight into the experiences of those with health anxiety. For example, medical condition or disorder words such as "dpdr" describe a significant state of distress or comorbid disorders. Disease keywords such as "fowleri" and "amyotrophic" describe fatal diseases. Words related to online resources are indicative of information sharing between group members. Using the CBT model to conceptualize word usage assisted in providing a map for further understanding the anxiety cycle and from where words and thoughts may emanate. The information related to the uniqueness of this corpus may provide clinicians, researchers, and other counseling professionals with a more enriched view of the experiences related to health anxiety and further serve Yalom's (2013) charge to enter the client's internal world.

Clinical Implications

The results of this study produced a list of medical-, symptom-, and anxiety-based language that appears consistent with previous theories of health anxiety. As such, the results

appear to support the validity of the subreddit in understanding the experiences of those with health anxiety. Clinicians may view keyword, collocates and consult the r/HealthAnxiety subreddit directly as a learning opportunity.

Clinicians may benefit from knowledge regarding symptom keywords such as "lightheadedness." Lightheadedness is observed as biproduct of increased anxiety and hyperventilation for those with health anxiety (Abramowitz & Deacon, 2004). Abramowitz & Deacon observe that the autonomic nervous system increases in activation due to elevated perception of perceived danger. As such, multiple physiological mechanisms are engaged to prepare the individual to engage in a fight or flight response. For clinician, a knowledge of keywords related to physiological arousal associated with health anxiety may provide additional information or opportunities for further assessment or treatment interventions as appropriate.

There may be some benefit in reading posts and replies regarding extremes of distress experienced, medical tests, and medical interventions used, and the struggle that many with health anxiety experience. Counselor educators could engage counselors-in-training with keyword results and viewing posts and comments on the subreddit r/HealthAnxiety as a learning opportunity for counselors-in-training.

Finally, counseling professionals could engage in advocacy within this forum and other mental health forums to remove predatory product advertisements, especially products that promote suicide attempts. Keywords for medications or supplements such as "nembutal" and "tanax" are used for euthanasia. According to Nepon et al. (2010), for individuals with a lifetime history of suicide attempts, over 70% had an anxiety disorder diagnosis. The raw corpus data reveal that these terms are largely related to companies attempting to market their products to the

health anxiety community. An advocacy opportunity exists for professional counselors to potentially volunteer as moderators of a subreddit to remove such listings.

Research Implications

To our knowledge, this study represents the first keyword and collocation analysis specific to health anxiety for monolingual English in online discourse. The results may serve as a baseline for comparison in further research. For example, studying linguistic changes during the COVID-19 pandemic may be useful to observe how a medical emergency impacts the language of those with health anxiety. In addition, concordance analysis or keywords in context (KWIC) analysis may provide additional information related to how keywords are used within the discussion board. Comparisons of these findings with other subreddits and support forums could potentially highlight differences in communication that are of note to clinicians who desire to know more about the experiences of those with health anxiety. Further, investigating the impact of such online forums on their users could shed light on their utility; for example, investigating the impact of reading and/or participating in such forums can help individuals to address their health concerns. Qualitative studies investigating the experience of participants in online forums may provide additional insights into the utility of forum participation.

Limitations

Generalization of results should be taken with care as the study corpus represents communication on a specific platform, in a specific language, and during a specific timeframe. The research conducted was in monolingual U.S. dialect English and, therefore, should not be generalized to other locales or languages. Online communication may vary from other mediums of communication, and as such, generalization across mediums is beyond the scope of this research project.

The reference corpus represents general web-based entries as identified by Google.com through the year 2012. Differences may emerge when comparing a study corpus to other reference corpora, and as such, a different set of keywords may be present. The study corpus is taken from the year 2019; communication may change as new words and expressions enter the lexicon. Proper nouns related to current events may produce a variation in keywords, especially related to the reference corpus understudy. As such, generalization should be taken with care and further replication may be warranted prior to making definitive conclusions related to the findings within this study. Finally, this study corpus is comprised of entries from Reddit.com, but health anxiety communication may vary across platforms such as Twitter and Facebook. However, generalization across platforms is beyond the scope of this research study.

Overall Conclusions

The keywords and collocates list of the study corpus provided insight into the uniqueness of health anxiety communication online. As Pennebaker et al. (2015) asserted, words reveal important information related to beliefs, thinking patterns, social relationships, personality, and fears. The keywords of this population correspond with a high-degree of medical-based language and associative. As the study of words evolves within the counseling profession there is an opportunity to enhance the understanding of the experience of those with various disorders. Doing so may have a myriad of effects such as enhancing treatment plans, training materials, and increasing empathy for those who suffer. Corpus-based research is in a nascent state for counseling researchers. As future studies employing similar methodologies, we anticipate that results will increase the counseling profession's comprehensive understanding of a variety of conditions.

References

Abramowitz, J.S., & Deacon, B. J. (2004). Severe health anxiety: Why it persists and how to treat it. *Comprehensive Therapy*, *30*(1), 44–49. https://doi.org/10.1007/s12019-004-0023-1

Abramowitz, J. S., Deacon, B. J., & Valentiner, D. P. (2007). The short health anxiety inventory: Psychometric properties and construct validity in a non-clinical sample. *Cognitive Therapy and Research*, *31*(6), 871–883. https://doi.org/10.1007/s10608-006-9058-1

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Author.

Anthony, L. (2020). AntConc (3.5.9) [Computer software]. Tokyo, Japan: Waseda University. https://www.laurenceanthony.net/software

Brezina, V. (2018). *Statistics in corpus linguistics: A practical guide*. Cambridge.

The COCA Corpus. (2021, June 12). https://www.english-corpora.org/coca/help/coca2020_overview.pdf

Dean, B. (2021, February 25). *Reddit usage and growth statistics: How many people in use Reddit in 2021?* https://backlinko.com/reddit-user

De Choudhury, M., & De, S. (2014). Mental health discourse on Reddit: Self-disclosure, social support, and anonymity. *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*, 71–80. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8075/8107

Egbert, J., & Biber, D. (2019). Incorporating text dispersion into keyword analyses. *Corpora*, *14*(1), 77–104. https://doi.org/10.3366/cor.2019.0162

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*(4)*,* 1149-1160. https://doi.org/10.3758/BRM.41.4.1149

Firth, J. R. (1968). *Selected papers of J.R. Firth 1952-1959* (F. R. Palmer, Ed.). Longman.

Furer, P., Walker, J. R., & Stein, M. B. (2007). *Treating health anxiety and fear of death: A practitioner's guide*. Springer. https://doi.org/10.1007/978-0-387-35145-2

Gabrielatos, C. (2018). *Keyness analysis: Nature, metrics and techniques*. In C. Taylor & A. Marchi (Eds.), *Corpora approaches to discourse: A critical review* (pp. 225–258). Routledge.

Geisler, J., & Dykeman, C. (2021, April 28). Word choice and collocates of Doka and Martin's grieving styles. *ArXiv.* https://doi.org/10.31234/osf.io/b3wf2

Greaves, M. M., & Dykeman, C. (2018). A corpus linguistic analysis of public Reddit blog posts on non-suicidal self-injury. *ArXiv*. https://doi.org/abs/1902.06689

Hardie, A. (2014, April 28). *Log ratio – an informal introduction.* http://cass.lancs.ac.uk/log-ratio-an-informal-introduction/

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

Jones, S. L., Hadjistavropoulos, H. D., & Gullickson, K. (2014). Understanding health anxiety following breast cancer diagnosis. *Psychology, Health & Medicine*, *19*(5), 525–535. https://doi.org/10.1080/13548506.2013.845300

LaGue, A., Cazares-Cervantes, A., Dykeman, C., Muzacz, A., & List, A. (2019, August 26). Research article length, design, and impact in counselor education: A worldwide Bayesian analysis. *ArXiv*. https://doi.org/10/31234/osf.io/qnsvm

Loper, E., & Bird, S. (2002). NLTK: The natural language toolkit. *ETMTNLP '02: Proceedings of the ACL-02 workshop on effective tools and methodologies for teaching natural language processing and computational linguistics*, *1*, 63-70. https://dl.acm.org/doi/proceedings/10.5555/1118108

Leech, G. (2002). The importance of reference corpora. *Hizkuntza-corpusak. Oraina eta geroa*, *10*(24/25), 1-11. https://www.uzei.eus/wp-content/uploads/2017/06/06-Geoffrey-LEECH.pdf

Marcus, D. K., Hughes, K. T., & Arnau, R. C. (2008). Health anxiety, rumination, and negative affect: A mediational analysis. *Journal of Psychosomatic Research*, *64*(5), 495–501. https://doi.org/10.1016/j.jpsychores.2008.02.004

Nepon, J., Belik, S.-L., Bolton, J., & Sareen, J. (2010). The relationship between anxiety disorders and suicide attempts: Findings from the National Epidemiologic Survey on Alcohol and Related Conditions. *Depression and Anxiety*, *4*(9), 791–798. https://doi.org/10.1002/da.20674

Pennebaker, J. W. (2013). *The secret life of pronouns: What our words say about us*. Bloomsbury Press.

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC 2015*. University of Texas at Austin. https://repositories.lib.utexas.edu/bitstream/handle/2152/31333/LIWC2015_LanguageManual.pdf?Sequence=3

Reeves, D., Blickem, C., Vassilev, I., Brooks, H., Kennedy, A., Richardson, G., & Rogers, A. (2014). The contribution of social networks to the health and self-management of patients

with long-term conditions: A longitudinal study. *PLOS One*, *9*(6), e98340. https://doi.org/10.1371/journal.pone.0098340

Rosnow, R. L., & Rosenthal, R. (2009). "Effect sizes for experimenting psychologists": Correction to Rosnow and Rosenthal (2003). *Canadian Journal of Experimental Psychology*, *63*(2), 123. https://doi.org/10.1037/a0015528

Salkovskis, P. M., & Warwick, H. M. C. (2001). Meaning, misinterpretations, and medicine: A cognitive-behavioral approach to understanding health anxiety and hypochondriasis. In V. Starcevic & D. R. Lipsitt (Eds.), *Hypochondriasis: Modern perspectives on an ancient malady* (pp. 202–222). Oxford University Press.

Semrush, 2021 (2022, February 7). Top 100: The most visited websites in the US. https://www.semrush.com/blog/most-visited-websites/

Shen Hanwen, J., & Rudzicz, F. (2017). Detecting anxiety on Reddit. *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology,* 58–65.

Starcevic, V., & Aboujaoude, E. (2015). Cyberchondria, cyberbullying, cybersuicide, cybersex: "New" psychopathologies for the 21st century? *World Psychiatry*, *14*(1), 97–100. https://doi.org/10.1002/wps.20195

Starcevic, V., & Berle, D. (2014). Cyberchondria: Towards a better understanding of excessive health-related Internet use. *Expert Review of Neurotherapeutics*, *13*(2), 205–213. https://doi.org/10.1586/ern.12.162

Tankovska, H. (2021a, May 3). *Distribution of Reddit app users in the United States as of March, 2021*. https://www.statista.com/statistics/1125159/reddit-us-app-users-age/

Tankovska, H. (2021b, April 21). *Regional distribution of desktop traffic to Reddit.com as of December 2020, by country.* https://www.statista.com/statistics/325144/reddit-global-active-user-distribution/

Tankovska, H. (2021c, May 3). *Percentage of U.S. adults who use Reddit as of February 2021, by gender.* https://www.statista.com/statistics/261765/share-of-us-internet-users-who-use-reddit-by-gender/

Tankovska, H. (2021d, June 11). *Percentage of U.S. adults who use Reddit as of February 2019, by ethnicity*. https://www.statista.com/statistics/261770/share-of-us-internet-users-who-use-reddit-by-ethnicity/

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, *29*(1), 24–54. https://doi.org/10.1177/0261927X09351676

Taylor, S., & Asmundson, G. J. G. (2004). *Treating health anxiety: A cognitive-behavioral approach.* The Guilford Press.

Tyrer, P., Cooper, S., Tyrer, H., Wang, D., & Bassett, P. (2019). Increase in the prevalence of health anxiety in medical clinics: Possible cyberchondria. *International Journal of Social Psychiatry*, *65*(7–8), 566–569. https://doi.org/10.1177/0020764019866231

Tyrer, P., & Tyrer, H. (2018). Health anxiety: Detection and treatment. *BJPsych Advances*, *24*(1), 66–72. Cambridge Core. https://doi.org/10.1192/bja.2017.5

Walker, & Furer, P. (2008). Interoceptive Exposure in the Treatment of Health Anxiety and Hypochondriasis. *Journal of Cognitive Psychotherapy*, *22*(4), 366–378. https://doi.org/10.1891/0889-8391.22.4.366

Warwick, H. M. C., & Salkovskis, P. M. (1990). Hypochondriasis. *Behaviour Research and Therapy*, *28*(2), 105–117. https://doi.org/10.1016/0005-7967(90)90023-C

Weisser, M. (2016). *Practical corpus linguistics: An introduction to corpus-based language analysis*. John Wiley & Sons.

Yalom, I. D. (2013). *The gift of therapy: An open letter to a new generation of therapists and their patients*. Harper Perennial.

Chapter 4: General Conclusions

This chapter summarizes the findings and implications of the two dissertation studies examining the linguistic attributes of health anxiety communication in an online forum. Central to the investigation is the premise that the words people use reveal important information related to their beliefs, thinking patterns, social relationships, personality, and fears (Pennebaker et al., 2015). The first study focused on the categories of word usage and summary characteristics for those with health anxiety. The second study focused on what those with health anxiety were discussing; that is, what particular words were used to make the communication on the online forum unique. Merging the two studies, information related to online communication may provide counseling professionals with a clearer picture of the internal experiences of those individuals on the support forum.

Summary of Manuscript 1

The first study aimed to identify the linguistic attributes of individuals who post on a health anxiety-related online forum. Reddit was chosen as the platform, and the sample was selected from all comments and replies from the subreddit r/HealthAnxiety for the year 2019. A full year was selected to control for possible variance due to seasonal changes. The year 2019 was selected as that year represents the most frequent communication prior to the COVID-19 pandemic that emerged in early 2020. The reference corpus selected was the web section of the Corpus of Contemporary American English (COCA; The COCA, 2021). This corpus consisted of a variety of blogs, political discourse, and other commentaries as identified by Google. The year of creation was 2012. Significant differences were found between the two corpora based on variables of interest. A synchronic corpus linguistic design was used to address the research questions.

Two python-coded scripts were used to acquire all the comments and replies from the Reddit database using the available API. The first script acquired the original posts, and the second script acquired the comments. Posts and comments from the calendar year 2019 were extracted and the two resulting text files were merged into one text file. Subsequently, the text file went through a preprocessing phase. During this phase, words that were not in monolingual U.S. English were converted to U.S. spelling (e.g., U.K. spelling of “colour” to U.S. spelling of “color”). Common web-based acronyms were converted to literal expressions (e.g., “lol” to “laugh out loud”). URLs with identifying data or personal links were removed from the corpus. Common misspellings were changed for software analysis (e.g., “ive” to “I've”).

The variables selected to measure were intended to build on a prior analysis of Shen and Rudzicz (2017). These researchers identified common variables across anxiety-specific subreddits. Additionally, broad linguistic category variables were captured along with biological processes such as body, ingest, and sexual. Descriptive and inferential statistics were gathered, and the results indicate distinct differences between the study and the reference corpus.

For the study corpus, there were 80,129 posts and comments and 5,461,459 tokens. Comparatively, the reference corpus had 98,796 entries for a total of 106,568,862 tokens. Descriptive attributes of the study corpus indicate high degrees of authenticity with an LIWC standardized score of 85.48 indicating elevated levels of spontaneous and unfiltered speech (Pennebaker Conglomerates, 2022). The results indicated a high degree of negative emotions and a low degree of positive emotions. The analytic scores showed lower degrees of analytic thought process within comments and replies. Details for descriptive statistics in comparison can be viewed in Table 2.1.

Inferential statistics indicated significant differences across variables between the two corpora. The log-likelihood scale and Bayes factor (BIC) tests were used to determine effect size and evidence again the null hypothesis across variables. Broad category variables, such as authenticity, tone, analytic, and clout, were not included in inferential statistics due to the proprietary calculation within the LIWC software. First-person singular, second-person, anxiety, sad, body, and health all exceeded the log-likelihood critical value for significance a $p < .01$. Bayes factor (BIC) indicated very strong evidence against H_0 across all variables. In contrast, first-person plural, third-person singular, positive emotion, anger, and sexual indicated lower use levels within the study corpus compared to the reference corpus.

The discussion compared the linguistic attributes with the health anxiety cycle as proposed by the cognitive-behavioral model of health anxiety. The results were interpreted as being in alignment with the model. Linguistic attributes may map to certain discrete steps within the cycle, adding additional information about how expressions of the individuals may be reflective of various points within the cycle. Ultimately, the results provided more information for counseling professionals to take into account when considering the experiences of those with health anxiety.

Summary of Manuscript 2

The purpose of the second study was to conduct a keyword and collocate analysis of health anxiety communication on a web-based internet forum. In this study, we were interested in what words were unique to the health anxiety corpus and what those words may indicate about communication within the online forum. Therefore, a keyword analysis, which identifies words that makes a text unique, was identified as the appropriate methodology.

The corpus was constructed as described within the first study, including scripting, compiling, and preprocessing steps. An additional preprocessing step was added to both corpora. Stop words as listed in the NLTK were removed from both corpora. The study used AntConc for the identification of the top keywords and collocates as well as for statical analysis (Anthony, 2020).

The results demonstrated significant differences between the two corpora for keywords. The total number of tokens in the study was 2,807,341, compared to the reference corpus at 56,483,561 tokens. A keyword list was created for both the study and reference corpus. The top 100 keywords all exceeded the minimum threshold for the log-likelihood scale for the critical value of 6.63. Hardie's (2014) log-ratio scale was used for effect size.

The top five types from the study corpus's list of the top 100 keywords were selected, and the term "health anxiety," to identify collocates. The results are listed in Table 3.3. The mutual information test was used to assess effect size. Collocates with a score about three are considered linguistically important (Hunston, 2002). Collocates provided additional information about the top five keywords and the term "health anxiety."

As the keywords were analyzed, certain patterns of categories of words emerged. These categories were identified, and words were label accordingly. Two doctoral level counselor educators evaluated the categories and found that they were appropriate. There was an initial 83% agreement of word categorization between the first author and the interraters. A discussion was held and specific words and word categories not in agreement were evaluated. After the discussion, the interrater agreement was 99.66%.

Within the discussion, word categories and specific words were compared with the health anxiety cycle as described in the cognitive behavioral model for health anxiety. Certain words

were evaluated regarding being reflective of the distress experienced by those with health anxiety. In addition, various terms were identified as potentially being reflective of reassurance seeking or reassuring giving. Ultimately, the keywords and categories provide additional information related to subreddit that may be helpful for counseling professionals across a variety of circumstances and professional duties. This may include clinicians wanting to know more about the experiences of those with the diagnosis, counselor educators seeking to augment learning activities for counselors-in-training, or counseling researchers who may be utilizing linguistic tools and techniques to provide further information and assistance to this population.

Limitations

There are several significant limitations. First, the data are comprised of monolingual English. As such, generalizing across international populations is not possible. Second, while some information related to the demographics of Reddit users exists, the exact population details of the users on r/HealthAnxiety are unknown, so generalization across demographic categories may not be feasible. Third, writings appear within a specific context—an online forum. Fourth, the reference corpus represents general web-based entries as identified by Google.com through the year 2012. This is the most recent year included in the COCA for online discourse (The COCA, 2021). When comparing discourse from 2019 to 2012, differences may emerge when comparing a study corpus to other reference corpora, and as such, a different set of keywords may be present, especially with the use of an alternative reference corpus. As such, generalization across timeframes may not be feasible and is beyond the scope of the present study. Individuals who suffer from health anxiety-related disorders may present differently from context to context.

Implications and Recommendations

To our knowledge, the two studies within this dissertation were the first studies to identify the linguistic attributes of health anxiety in an online context. Two findings from the data analysis are the high degree of self-focus and negative affect for those with health anxiety. While previous research has observed a high degree of self-focus within anxiety disorders, health-anxiety specific self-focus has not, to our knowledge, been an area of study. The elevated use of self-focus supports the notion that increased self-focus may be a transdiagnostic across a myriad of mental health disorders (Ingram, 1990). As such, a possible recommendation is for clinicians to increase identification and the implications related to elevated levels of self-focus. Counselor educators may consider augmenting training materials with identification and interventions that address high-levels of self-focus—for example, actively tuning in to clients' self-focused communications or descriptions of their experiences. Counseling researchers could investigate whether decreasing self-focus, specific to health anxiety, has an impact on overall symptoms for health anxiety. Additionally, future research could investigate whether linguistic elements of self-focus diminish over the course of treatment and is indicative of recovery from health anxiety.

Negative words usage was another finding in the study corpus. This supports previous research for health anxiety regarding the association of negative affect and health anxiety (Marcus et al., 2008; Mor & Winquist, 2002). Recommendations for counselors include identifying and using interventions appropriate to addressing negative cognitions when treating those with health anxiety—for example, applying interventions specific to addressing catastrophizing and emotional reasoning. Relatedly, counselor educators could consider the findings of the studies contained herein as learning opportunities. Counselor educators may

augmenting training materials specifically for addressing negative cognitions, such as cognitive distortion identification skills and remediation as proposed in CBT. In terms of research, counseling researchers could investigate whether, linguistically, negative language decreases over the course of treatment as a marker for overall improvement of the client.

Another finding is that redditors communicate authentically about their mental health experiences on the Reddit platform. This finding supports previous research (De Choudhury & De, 2014) and implies that redditors write authentically about their mental health experiences on mental health-specific subreddits. While it may be difficult based on the results of this research to generalize across all mental health subreddits, for the study corpus, the authenticity scores were elevated. Per this finding, it is recommended that counselors consider viewing comments and posts as a learning opportunity as they build competencies in working with health anxiety. Posts and comments can capture acute episodes of anxiety that may not be present when the client is in session. Viewing these posts and comments may provide additional insight into the experiences of those with health anxiety, and perhaps increase accurate empathy for those who suffer with health anxiety symptoms. Relatedly, it is recommended that counselor educators consider evaluating whether reading posts and comments of mental health discussion boards for counselors-in-training would be appropriate for their curriculum.

Further, it is recommended that counseling researchers evaluate other mental health subreddits to determine authenticity levels, and additional research into possible variance may be helpful in increasing the understanding of authentic communication across social media platforms. Research may be conducted to compare authenticity within a counseling context versus an online context, and investigation into the determining factors for possible variance may also be helpful for clinicians. Researchers utilizing qualitative methodologies may investigate the

process or experience of redditors expressing acute distress on health anxiety forums, especially in relation to key markers such as “atm,” an abbreviation for “at the moment.” Future research may investigate linguistic alterations between years 2019 and 2020 respectively per the COVID-19 global pandemic. Also, researchers may investigate linguistic changes over time for redditors on r/HealthAnxiety. This may provide clinician with relevant data regarding whether forum participation may be beneficial or potentially harmful for their clients.

Health anxiety is often underdiagnosed (Tyrer & Tyrer, 2018). The keyword list describes symptoms that may not be directly associated with health anxiety. For example, the keyword “lightheadedness” which appears in the top five of the top 100 keywords with health anxiety. “Lightheadedness” is not a description for illness anxiety disorder or somatoform disorder but has been noted within the literature as a biproduct of an anxiety response which may be present in those with health anxiety (Abramowitz & Deacon, 2004). It is recommended that clinicians view the keyword list as a learning opportunity to identify potential symptoms for suggestive of health anxiety. A knowledge of lesser known symptoms may assist clinicians in identifying markers in symptom description which may serve as a prompt for further assessment for clients.

Based on the findings of these studies, a profile of those with health anxiety emerged within the context of the study corpus. The redditors present as authentic; lower in clout; self-focused; high in negative affect; low in positive emotion; low in anger; high in words pertaining to body, health, and ingest; and low in sexual words. Additionally, the keyword results suggest a uniqueness of language including using language pertaining to medical conditions and diseases, medication and supplements, and medical tests. The implication of the profile suggests that linguistic tools may be useful in providing profiles, at a group level, of varying pathologies.

It is recommended that counseling researchers investigate profile construction and how they may vary across contexts—for example, comparing social media profiles versus in-person profiles. Also, counseling researchers should investigate how such profiles may inform clinical practice. Counselors may find knowledge of linguistic profiles as informational. Such knowledge may increase attention to markers in the language of their clients. For example, monitoring for specific medical nomenclature, especially from those individuals who do not work in the medical field. Counselors may also ask general questions about health and health-related anxiety throughout treatment and note markers that may indicate that further assessment is appropriate. Counselor educators may consider viewing results from profile construction to assist counselors-in-training with diagnosis and treatment based on the findings from future research. Adding profile analysis to training courses could provide counseling researchers an opportunity to investigate the efficacy of considering profiles in accurate diagnosis in counseling programs.

At an enterprise or platform level, real-time linguistic analysis can be programmed to send alerts for posts or comments that indicate that a user is at risk. For example, De Choundry et al. (2013) called for the development of tools based on real-time markers for post-partum depression. In terms of implications, a new sphere of counseling may be developed, that of enterprise-level counseling. Counselors could partner with larger platforms to develop tools and interventions for at-risk populations. It is recommended that future research investigate the efficacy of a platform and counseling partnership and assess the effectiveness of such a partnership.

A view of the keyword in context revealed that posts seem to originate from retailers seeking to sell euthanasia products to the members of this message board, thus conceivably posing the risk of promoting suicide. The implication is that marketers are offering their products

to at-risk populations. It is recommended that counselors engage with message boards and potentially act as moderators to remove postings that offer euthanasia products to potentially at-risk populations. Clinicians should consider discussing with their clients who participate in such forums the potential risk factors impacting their wellness.

A significant potential for clinical application exists for the development of assessment tools. There is a significant need for identification of health anxiety as it is often underdiagnosed Tyrer & Tyrer, 2019). As Coppersmith et al. (2014) reported the ability to distinguish disorders by analyzing language within twitter messages, tools may be developed to allow for language analysis that provides information pertaining to diagnosis. Researchers should continue to identify linguistic variables across pathologies and in a variety of contexts. As a knowledge-base of linguistic attributes related to mental health diagnosis is constructed, researchers may develop and validate measures that analyze language for use in assessment. For instance, a software application may be created to analyze a sample of writing from a client. Based on the linguistic analysis, diagnostic impressions, and other information relevant to treatment could be delivered in a report for the clinician.

Conclusion

The two studies in this dissertation sought to identify the linguistic characteristics of individuals with health anxiety. To do so, we engaged in a naturalistic observational approach using data from a health anxiety-dedicated forum for those who sought to overcome their anxiety. The results produced, to our knowledge, the first analysis of the linguistic attributes specific to health anxiety communication in an online context. In combination, the two studies supported previous research related to health anxiety and constructed a linguistic profile of health anxiety communication in an online context. It is hoped that the results from these studies

will be viewed as a baseline from which future research will build. Future studies may reveal whether addressing linguistic attributes directly is helpful for clinicians, or whether assessment of linguistic changes throughout the course of treatment is indicative of symptom amelioration. Also, we anticipate that knowledge generated by these two studies will increase understanding of those with health anxiety, their linguistic attributes, and markers of distress, and that it will assist counselors in treatment, training, and development of future research.

Bibliography

Abramowitz, J.S., & Deacon, B. J. (2004). Severe health anxiety: Why it persists and how to treat it. *Comprehensive Therapy*, *30*(1), 44–49. https://doi.org/10.1007/s12019-004-0023-1

Abramowitz, J. S., Deacon, B. J., & Valentiner, D. P. (2007). The short health anxiety inventory: Psychometric properties and construct validity in a non-clinical sample. *Cognitive Therapy and Research*, *31*(6), 871–883. https://doi.org/10.1007/s10608-006-9058-1

Al-Mosaiwi, M., & Johnstone, T. (2018). In an absolute state: Elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation. *Clinical Psychological Science*, *6*(4), 529–542. https://doi.org/10.1177/2167702617747074

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Author.

Asmundson, G. J. G., & Taylor, S. (2020). How health anxiety influences responses to viral outbreaks like COVID-19: What all decision-makers, health authorities, and health care professionals need to know. *Journal of Anxiety Disorders*, *71*, Article 102211. https://doi.org/10.1016/j.janxdis.2020.102211

Anthony, L. (2020). AntConc (3.5.9) [Computer software]. Tokyo, Japan: Waseda University. https://www.laurenceanthony.net/software

Brezina, V. (2018). *Statistics in corpus linguistics: A practical guide*. Cambridge.

Bucci, W., & Freedman, N. (1981). The language of depression. *Bulletin of the Menninger Clinic*, *45*(4), 334–358. https://www.proquest.com/openview/d804439a2c70467603bbdf0c20a3f31a/1?pq-origsite=gscholar&cbl=1818298

Choudhury, M. D., Counts, S., & Horvitz, E. (2013). Major life changes and behavioral markers in social media. Proceedings of the 2013 Conference on Computer Supported Cooperative Work - CSCW 13. https://dx.doi.org/10.1145/2441776.2441937

Chung, C., & Pennebaker, J. W. (2007). The psychological functions of function words. In K. Fielder (Ed.), *Social communication* (pp. 343–359). Psychology Press.

The COCA Corpus. (2021, June 12). https://www.english-corpora.org/coca/help/coca2020_overview.pdf

Coppersmith, G., Dredze, M., & Harman, C. (2014). Quantifying mental health signals in Twitter. Proceedings of the *Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 51–60. https://doi.org/10.3115/v1/W14-3207

De Choudhury, M., Counts, S., & Horvitz, E. (2013). Major life changes and behavioral markers in social media: Case of childbirth. *CSCW '13: Proceedings of the 2013 conference Computer Supported Cooperative Work*, 1431–1442. https://doi.org/10.1145/2441776.2441937

De Choudhury, M., & De, S. (2014). Mental health discourse on Reddit: Self-disclosure, social support, and anonymity. *Proceedings of the International AAAI Conference on Web and Social Media*, *8*(1). https://ojs.aaai.org/index.php/ICWSM/article/view/14526

Dean, B. (2021, February 25). *Reddit usage and growth statistics: How many people in use Reddit in 2021?* https://backlinko.com/reddit-user

Egbert, J., & Biber, D. (2019). Incorporating text dispersion into keyword analyses. *Corpora*, *14*(1), 77–104. https://doi.org/10.3366/cor.2019.0162

Eisner, M. D., Blanc, P. D., Yelin, E. H., Katz, P. P., Sanchez, G., Iribarren, C., & Omachi, T. A. (2010). Influence of anxiety on health outcomes in COPD. *Thorax*, *65*(3), 229–234. https://doi.org/10.1136/thx.2009.126201

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*(4)*,* 1149-1160. https://doi.org/10.3758/BRM.41.4.1149

Fink, P., Ørnbøl, E., & Christensen, K. S. (2010). The outcome of health anxiety in primary care. a two-year follow-up study on health care costs and self-rated health. *PLOS One*, *5*(3), e9873. https://doi.org/10.1371/journal.pone.0009873

Firth, J. R. (1968). *Selected papers of J.R. Firth 1952-1959* (F. R. Palmer, Ed.). Longman.

Fischer-Homberger, E. (1972). Hypochondriasis of the eighteenth century—neurosis of the present century. *Bulletin of the History of Medicine*, *46*(4), 391–401.

Freud, S. (1960). *Psychopathology of everyday life*. E. Benn.

Furer, P., Walker, J. R., & Stein, M. B. (2007). *Treating health anxiety and fear of death: A practitioner's guide*. Springer. https://doi.org/10.1007/978-0-387-35145-2

Gabrielatos, C. (2018). *Keyness analysis: Nature, metrics and techniques*. In C. Taylor & A. Marchi (Eds.), *Corpora approaches to discourse: A critical review* (pp. 225–258). Routledge.

Gaur, M., Kursuncu, U., Alambo, A., Sheth, A., Daniulaityte, R., Thirunarayan, K., & Pathak, J. (2018). "Let me tell you about your mental health!": Contextualized classification of Reddit posts to DSM-5 for web-based intervention. *CIKM '18: Proceedings of the 27th*

ACM International Conference on Information and Knowledge Management, 753–762. https://doi.org/10.1145/3269206.3271732

Geisler, J., & Dykeman, C. (2021, April 28). Word choice and collocates of Doka and Martin's grieving styles. *ArXiv.* https://doi.org/10.31234/osf.io/b3wf2

Greaves, M. M., & Dykeman, C. (2018). A corpus linguistic analysis of public Reddit blog posts on non-suicidal self-injury. *ArXiv*. https://doi.org/abs/1902.06689

Hadjistavropoulos, H. D., Janzen, J. A., Kehler, M. D., Leclerc, J. A., Sharpe, D., & Bourgault-Fagnou, M. D. (2012). Core cognitions related to health anxiety in self-reported medical and non-medical samples. *Journal of Behavioral Medicine*, *35*(2), 167–178. https://doi.org/10.1007/s10865-011-9339-3

Hardie, A. (2014, April 28). *Log ratio – an informal introduction.* http://cass.lancs.ac.uk/log-ratio-an-informal-introduction/

Hardie, A., McEnery, T., & Baker, P. (2006). *A glossary of corpus linguistics*. Edinburgh: Edinburgh University Press.

Hedman, E., Andersson, G., Andersson, E., Ljótsson, B., Rück, C., Asmundson, G., & Lindefors, N. (2011). Internet-based cognitive–behavioural therapy for severe health anxiety: Randomised controlled trial. British Journal of Psychiatry, 198(3), 230-236. doi:10.1192/bjp.bp.110.086843

Holmes, E. A., O'Connor, R. C., Perry, V. H., Tracey, I., Wessely, S., Arseneault, L., Ballard, C., Christensen, H., Silver, R. C., Everall, I., Ford, T., John, A., Kabir, T., King, K., Madan, I., Michie, S., Przybylski, A. K., Shafran, R., Sweeney, A., … Bullmore, E. (2020). Multidisciplinary research priorities for the COVID-19 pandemic: A call for

action for mental health science. *Lancet Psychiatry*, *7*(6), 547–560. https://doi.org/10.1016/S2215-0366(20)30168-1

Holtzman, N. S., Tackman, A. M., Carey, A. L., Brucks, M. S., Küfner, A. C. P., Deters, F. G., Back, M. D., Donnellan, M. B., Pennebaker, J. W., Sherman, R. A., & Mehl, M. R. (2019). Linguistic markers of grandiose narcissism: A LIWC analysis of 15 samples. *Journal of Language and Social Psychology*, *38*(5–6), 773–786. https://doi.org/10.1177/0261927X19871084

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

Ingram, R. E. (1990). Self-focused attention in clinical disorders: Review and a conceptual model. *Psychological Bulletin*, *107*(2), 156–176. https://doi.org/10.1037/0033-2909.107.2.156

Jones, S. L., Hadjistavropoulos, H. D., & Gullickson, K. (2014). Understanding health anxiety following breast cancer diagnosis. *Psychology, Health & Medicine*, *19*(5), 525–535. https://doi.org/10.1080/13548506.2013.845300

Kern, M. L., Park, G., Eichstaedt, J. C., Schwartz, H. A., Sap, M., Smith, L. K., & Ungar, L. H. (2016). Gaining insights from social media language: Methodologies and challenges. *Psychological Methods*, *21*(4), 507–525. https://doi.org/10.1037/met0000091

LaGue, A., Cazares-Cervantes, A., Dykeman, C., Muzacz, A., & List, A. (2019, August 26). Research article length, design, and impact in counselor education: A worldwide Bayesian analysis. *ArXiv*. https://doi.org/10/31234/osf.io/qnsvm

Leech, G. (2002). The importance of reference corpora. *Hizkuntza-corpusak. Oraina eta geroa*, *10*(24/25), 1-11. https://www.uzei.eus/wp-content/uploads/2017/06/06-Geoffrey-LEECH.pdf

Loper, E., & Bird, S. (2002). NLTK: The natural language toolkit. *ETMTNLP '02: Proceedings of the ACL-02 workshop on effective tools and methodologies for teaching natural language processing and computational linguistics*, *1*, 63-70. https://dl.acm.org/doi/proceedings/10.5555/1118108

Marcus, D. K., Hughes, K. T., & Arnau, R. C. (2008). Health anxiety, rumination, and negative affect: A mediational analysis. *Journal of Psychosomatic Research*, *64*(5), 495–501. https://doi.org/10.1016/j.jpsychores.2008.02.004

Mathieson, F., Jordan, J., Carter, J. D., & Stubbe, M. (2016). Nailing down metaphors in CBT: Definition, identification and frequency. *Behavioural and Cognitive Psychotherapy*, *44*(2), 236–248. https://doi.org/10.1017/S1352465815000156

Mor, N., & Winquist, J. (2002). Self-focused attention and negative affect: A meta-analysis. *Psychological Bulletin*, *128*(4), 638–662. https://doi.org/10.1037//0033-2909.128.4.638

Newby, J. M., Smith, J., Uppal, S., Mason, E., Mahoney, A. E. J., & Andrews, G. (2018). Internet-based cognitive behavioral therapy versus psychoeducation control for illness anxiety disorder and somatic symptom disorder: A randomized controlled trial. *Journal of Consulting and Clinical Psychology*, *86*(1), 89–98. https://doi.org/10.1037/ccp0000248

O'Bryan, E. M., McLeish, A. C., & Johnson, A. L. (2017). The role of emotion reactivity in health anxiety. *Behavior Modification*, *41*(6), 829–845. https://doi.org/10.1177/0145445517719398

Park, A., & Conway, M. (2017). Longitudinal changes in psychological states in online health community members: Understanding the long-term effects of participating in an online

depression community. *Journal of Medical Internet Research*, *19*(3), e71. https://doi.org/10.2196/jmir.6826

Park, M., Cha, C., & Cha, M. (2012). Depressive moods of users portrayed in twitter. *Proceedings of the ACM SIGKDD Workshop on Healthcare Informatics (HI-KDD)2012*, 1–8.

Pennebaker Conglomerates. (2021, October 14). Interpreting LIWC Output. https://liwc.wpengine.com/interpreting-liwc-output/

Pennebaker Conglomerates. (2022, February 2). *LIWC analysis*. https://www.liwc.app/help/liwc

Pennebaker, J. W. (2013). *The secret life of pronouns: What our words say about us*. Bloomsbury Press.

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC 2015*. University of Texas at Austin. https://repositories.lib.utexas.edu/bitstream/handle/2152/31333/LIWC2015_LanguageManual.pdf?Sequence=3

Qiu, L., Lin, H., Ramsay, J., & Yang, F. (2012). You are what you tweet: Personality expression and perception on Twitter. *Journal of Research in Personality*, *46*(6), 710–718. https://doi.org/10.1016/j.jrp.2012.08.008

R/Health Anxiety (2021, June 20). R/HealthAnxiety. https://www.reddit.com/r/HealthAnxiety/

Reeves, D., Blickem, C., Vassilev, I., Brooks, H., Kennedy, A., Richardson, G., & Rogers, A. (2014). The contribution of social networks to the health and self-management of patients with long-term conditions: A longitudinal study. *PLOS One*, *9*(6), e98340. https://doi.org/10.1371/journal.pone.0098340

Ronen, T. (2011). Using metaphors in therapy. In T. Ronen (Ed.), *The positive power of imagery* (pp. 123–135. Publisher. https://doi.org/10.1002/9780470979976.ch8

Rosnow, R. L., & Rosenthal, R. (2009). "Effect sizes for experimenting psychologists": Correction to Rosnow and Rosenthal (2003). *Canadian Journal of Experimental Psychology*, *63*(2), 123–123. https://doi.org/10.1037/a0015528

Rude, S., Gortner, E.-M., & Pennebaker, J. (2004). Language use of depressed and depression-vulnerable college students. *Cognition and Emotion*, *18*(8), 1121–1133. https://doi.org/10.1080/02699930441000030

Saha, K., Torous, J., Caine, E. D., & De Choudhury, M. (2020). Psychosocial effects of the COVID-19 pandemic: Large-scale quasi-experimental study on social media. *Journal of Medical Internet Research*, *22*(11), e22600. https://doi.org/10.2196/22600

Salkovskis, P. M., & Warwick, H. M. C. (2001). Meaning, misinterpretations, and medicine: A cognitive-behavioral approach to understanding health anxiety and hypochondriasis. In V. Starcevic & D. R. Lipsitt (Eds.), *Hypochondriasis: Modern perspectives on an ancient malady* (pp. 202–222). Oxford University Press.

Schmidt, N. B., Joiner, T. E., Staab, J. P., & Williams, F. M. (2003). Health perceptions and anxiety sensitivity in patients with panic disorder. *Journal of Psychopathology and Behavioral Assessment*, *25*(3), 139–145. https://doi.org/10.1023/A:1023520605624

Semrush, 2021 (2022, February 7). Top 100: The most visited websites in the US. https://www.semrush.com/blog/most-visited-websites/

Shen Hanwen, J., & Rudzicz, F. (2017). Detecting anxiety on Reddit. *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology, Vancouver Canada*, 58–65.

Simmons, R. A., Chambless, D. L., & Gordon, P. C. (2008). How do hostile and emotionally overinvolved relatives view relationships?: What relatives' pronoun use tells us. *Family Process*, *47*(3), 405–419. https://doi.org/10.1111/j.1545-5300.2008.00261.x

Sonnenschein, A. R., Hofmann, S. G., Ziegelmayer, T., & Lutz, W. (2018). Linguistic analysis of patients with mood and anxiety disorders during cognitive behavioral therapy. *Cognitive Behaviour Therapy*, *47*(4), 315–327. https://doi.org/10.1080/16506073.2017.1419505

Starcevic, V., & Aboujaoude, E. (2015). Cyberchondria, cyberbullying, cybersuicide, cybersex: "New" psychopathologies for the 21st century? *World Psychiatry*, *14*(1), 97–100. https://doi.org/10.1002/wps.20195

Starcevic, V., & Berle, D. (2014). Cyberchondria: Towards a better understanding of excessive health-related Internet use. *Expert Review of Neurotherapeutics*, *13*(2), 205–213. https://doi.org/10.1586/ern.12.162

Tankovska, H. (2021a, May 3). *Distribution of Reddit app users in the United States as of March, 2021*. https://www.statista.com/statistics/1125159/reddit-us-app-users-age/

Tankovska, H. (2021b, April 21). *Regional distribution of desktop traffic to Reddit.com as of December 2020, by country.* https://www.statista.com/statistics/325144/reddit-global-active-user-distribution/

Tankovska, H. (2021c, May 3). *Percentage of U.S. adults who use Reddit as of February 2021, by gender.* https://www.statista.com/statistics/261765/share-of-us-internet-users-who-use-reddit-by-gender/

Tankovska, H. (2021d, June 11). *Percentage of U.S. adults who use Reddit as of February 2019, by ethnicity*. https://www.statista.com/statistics/261770/share-of-us-internet-users-who-use-reddit-by-ethnicity/

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, *29*(1), 24–54. https://doi.org/10.1177/0261927X09351676

Taylor, S., & Asmundson, G. J. G. (2004). *Treating health anxiety: A cognitive-behavioral approach.* The Guilford Press.

Tyrer, P., & Tyrer, H. (2018). Health anxiety: Detection and treatment. *BJPsych Advances*, *24*(1), 66–72. Cambridge Core. https://doi.org/10.1192/bja.2017.5

Tyrer, P., Cooper, S., Tyrer, H., Wang, D., & Bassett, P. (2019). Increase in the prevalence of health anxiety in medical clinics: Possible cyberchondria. *International Journal of Social Psychiatry*, *65*(7–8), 566–569. https://doi.org/10.1177/0020764019866231

Walker, & Furer, P. (2008). Interoceptive Exposure in the Treatment of Health Anxiety and Hypochondriasis. *Journal of Cognitive Psychotherapy*, *22*(4), 366–378. https://doi.org/10.1891/0889-8391.22.4.366

Warwick, H. M. C., & Salkovskis, P. M. (1990). Hypochondriasis. *Behaviour Research and Therapy*, *28*(2), 105–117. https://doi.org/10.1016/0005-7967(90)90023-C

Weisser, M. (2016). *Practical corpus linguistics: An introduction to corpus-based language analysis*. John Wiley & Sons.

White, M., & Epston, D. (1990). *Narrative means to therapeutic ends*. WW Norton.

Wilson, A. (2013). Embracing Bayes factors for key item analysis in corpus linguistics. In M. Bieswanger, & A. Koll-Stobbe (Eds.), *New approaches to the study of linguistic variability* (pp. 3-11). (Language Competence and Language Awareness in Europe; Vol. 4). Peter Lang.

Yalom, I. D. (2013). *The gift of therapy: An open letter to a new generation of therapists and their patients*. Harper Perennial.

Appendix

IRB exemption -

On Jul 27, 2021, at 1:35 PM, IRB <irb@oregonstate.edu> wrote:

I apologize for the delay Christopher. If you will only be accessing data that are publicly available where there is no reasonable expectations of privacy—similar to what you described in your email—your study would not be considered research with human subjects and you would not need to apply for this office. However, if you need an official determination on a specific activity, you would need to apply to our online system. If this tis the case, le me know and I will get you moving on that process. Thanks for the questions and good luck! It sounds like you have some interesting projects lined up.

Best,
Nat Krancus, MA, CIP
Senior Analyst
Human Research Protection Program
Office of Research Integrity | **Oregon State University**
nat.krancus@oregonstate.edu

From: Christopher G McBride <mcbrichr@oregonstate.edu>
Sent: Tuesday, July 20, 2021 8:32 AM
To: IRB <irb@oregonstate.edu>
Subject: Re: Questions about exemption for two proposed studies

Hello,

I wanted to follow up on my email I sent last week (please see below). I wanted to check-in to determine if you required any further information or if there were any questions, etc. Any assistance would be much appreciated. Many thanks!

-Chris McBride

On Jul 13, 2021, at 9:43 AM, Christopher G McBride <mcbrichr@oregonstate.edu> wrote:

Hello!

Happy Tuesday. My name is Chris McBride. I'm a Ph.D. candidate within the Counseling unit at OSU. My advisor (and PI) is Kok-Mun Ng, Ph.D. Per the requirements of my dissertation, I'm proposing two research studies. I'm writing to determine if these studies would qualify for an IRB exemption.

The studies utilize public comments and replies taken from Reddit.com. Comments and replies are readable to any user on the internet who goes toReddit.com (thus, no privacy is implied for those who post on Reddit.com).

For further information.

My proposed sample is:

- One calendar year of comments and replies taken from the subreddit r/HealthAnxiety (year of 2019)
- Blog posts as identified by Google.com extracted from the Corpus of Contemporary American English (COCA) (a large collection of available writings)

For study 1, my research questions are:

1. What is the level of use of linguistic processes in online posts about health anxiety?
2. What is the pattern of use of psychological processes in online posts about health anxiety?
3. What is the score of summary variables about health anxiety?
4. What is the level of use of broad psycholinguistic processes in online posts about health anxiety compared to a reference corpus?
5. What is the pattern of use of linguistic processes variables in online posts about health anxiety compared to a reference corpus?
6. What is the pattern of use of psychological processes in online posts about health anxiety compared to a reference corpus?

For study 2, my research questions are:

1. What words distinguish online posts about health anxiety from online posts in general?
2. What words distinguish general online posts from online posts about health anxiety?
3. What are the most common collocations of the strongest keywords of online posts about health anxiety?
4. What are the most common collocations of the term “health anxiety” in online posts about health anxiety?

I am specifically looking at the broad use of language on the subreddit, and not specific user posts. Usernames will not be used in the analysis of this data.

My questions are:

1. Do these two studies qualify for an IRB exemption?
2. What information would be important for the IRB to have, within the IRB application, to clarify that these studies would be eligible for an exemption?

I would be happy to discuss this in further detail via phone conversation or Zoom call. I attempted to call during office hours yesterday, but the voicemail indicated that I should email my questions to irb@oregonstate.edu.

Many thanks in advance for any information you may provide. I hope you have a wonderful day!

Kind Regards,

Christopher McBride
Pronouns: he/his/him
Ph.D. Candidate
Counseling
OSU