PATTERNS OF HEALTH: A CORPUS ANALYSIS OF HEALTH INFORMATION AND MESSAGING

by

KATHERINE IRELAND KUIPER

(Under the Direction of Bill Kretzschmar)

ABSTRACT

This work analyzes public health communication in English across different contexts and time. The datasets involved include the *Royal Society Corpus* (Fischer et al. 2020), a corpus of US and UK press on e-cigarettes, and a corpus of CDC, WHO, and NHS Tweets during 2020. Each chapter presents clear findings displaying micro and macro-level changes in linguistic patterns and communication on notable and important topics and genres. This research further demonstrates the utility of multiple methods and incorporates R and Python programming for the compilation and analysis of the data. Finally, the relevance and applications of corpus linguistics for understanding ongoing concerns, including public health issues, are highlighted.

INDEX WORDS: corpus linguistics, public health communication, R Programming, press discourses, COVID-19, Twitter

PATTERNS OF HEALTH: A CORPUS ANALYSIS OF HEALTH INFORMATION AND MESSAGING

by

KATHERINE IRELAND KUIPER

M.A., University of Georgia, 2022

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2022

PATTERNS OF HEALTH: A CORPUS ANALYSIS OF HEALTH INFORMATION AND MESSAGING

by

KATHERINE IRELAND KUIPER

Major Professor:	Bill Kretzschmar
Committee:	Peggy Renwick
	Chad Howe

Electronic Version Approved:

Ron Walcott
Vice Provost for Graduate Education and Dean of the Graduate School
The University of Georgia
May 2022

DEDICATION

This work is dedicated to my mother and father.

ACKNOWLEDGEMENTS

It is with much gratitude that I acknowledge the encouragement, kindness, and support of my friends, family, and colleagues. I especially want to thank my advisor Bill Kretzschmar for his ongoing inspiration, help, and assistance. This work would not have been possible without his teaching, encouragement, and feedback. Many thanks to my committee members, Peggy Renwick and Chad Howe, who have provided invaluable comments and advice, and to the Linguistics department, to John Hale, Elliott Kuecker, Allison Burkette, and Amy Smoler. Thanks to my fellow graduate students and friends for their cheer and assistance: Mike and Rachel Olsen, Lisa Lipani, Camila Lívio, Joey Stanley, Renee Buesking, Melissa Gomes, Lorraine Van Hersch, Mary Ball Markow, Keiko Bridwell, Mathilde Daussy-Renaudin, Victor Renaudin, Donnie Dunagan, Jonathan Jones, and many others.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

> "Disease is a fundamental aspect of the human condition…[it] is something men and women feel. It is something in our bodies- but also in our minds…Disease demands explanation; we think about it and we think with it… Disease has always been a social and **linguistic** as well as biological entity." (Rosenberg 2007: vii, emphasis my own).

The primary goal of this dissertation is to uncover and examine linguistic devices and patterns using corpus-based methods on health-related communication datasets, of importance for both language research and public health concerns. Understandings of disease, health technology, and health-related decisions are informed through and by language patterns. Linguistic patterns in health-related messaging are of utmost importance and of special relevance today. By probing issues of public health and communication in important datasets with respect to both researchers, government officials, and members of the public alike, this work endeavors to answer the following research questions:

> What are the varied ways that lexico-grammatical patterns, multiword units, and different genres of English express issues of public health and health technology over time?
>
> What does the corpus-based study of variation in English reveal about the way different groups understand and discuss health and disease-related information?

Each dataset in this dissertation provides a different genre of language use with regard to public health, and the analytical outcome changes depending on the dataset. The fundamental theme is the same: a careful consideration of how health, health problems, and health-related technology is

expressed, discussed, and understood through language patterns and discourses. This is analyzed over time, beginning with the Early Modern English period, before public health as we know it today was widely understood or recognized (Berridge 2014). The datasets include the Royal Society corpus (RSC) (Fischer et al. 2020), a corpus of the *Philosophical Transactions of the Royal Society of London* (*Phil Trans*), the corpora of UK and US news on e-cigarettes from a 10 year time span (Kuiper 2020), and a corpus of NHS, WHO, and CDC tweets on COVID-19 (Kuiper forthcoming).

The first analysis chapter focuses on changes in discourses and linguistic patterns beginning in the Early Modern English period through analyzing the *Phil Trans*. Keywords and multiword units in the RSC change over the span of the corpus and reveal specific linguistic patterns vital to the development, communication and dissemination of science and the promotion of public health more broadly. Discourses of health and medicine are examined through keywords, including *health, disease(s),* and *inoculation,* integrated with collocational and concordance (KWIC) analysis. This work highlights the importance of corpus-based methods when dealing with trends across four hundred years of data and scholarly communication. The impact of modern science and changing scientific knowledge is exposed on medical and cultural ideas with discourses highlighting movement toward modern public health efforts. As new scientific ideas and theories were proposed and received, they were applied to the specific sciences of anatomy and medicine.

Chapter Five examines another notable dataset with regard to public health and cultural discourses: corpora of US and UK press coverage of e-cigarettes. These corpora include data from January 2010 through March 2020, with approximately 4 million tokens in total. This chapter highlights prevailing differences between media coverage in the US and the UK over time, due in part to differences in health-care systems. E-cigarettes are an especially interesting topic due to their connection with the tobacco industry and historical changes; they are also the most heavily used tobacco-substitute product by youth since 2014 in the US, with similar trends

present in the UK (Walley et al. 2019; Royal Society Report). The analysis of media coverage of e-cigarettes is essential due to the impact and power of press on the wider public; they provide an integral "source of societal information regarding scientific, technological, and medical developments" (2017). Examination of these corpora reflects the choices and framing of press sources across places and time. This chapter employs multiple corpus-based methods by evaluating collocations and frequencies of single and multiword units, along with bigrams. Each of these methods provide distinct advantages for the consideration of cultural issues in language patterns (Hunston 2010; Römer 2010; Scott and Tribble 2010). Findings include different themes of concerns between the US and UK press and more broadly a lack of clearly positive or negative language about e-cigarettes. US accounts generally trend toward emphasizing US-internal affairs, underage users, and economic interests. The UK press is more attentive to research and worldwide concerns.

Chapter Six analyzes a very recent dataset, utilizing all tweets from the official accounts of the National Health Service (NHS), the Centers for Disease Control (CDC), and the World Health Organization (WHO) in 2020 during the COVID-19 pandemic. Twitter data across these governmental organizations differs significantly over time and between entities, and the specific lexico-grammatical devices also vary including choice of verbs, passivization, and intensification. This chapter underscores the necessity of careful communication with regard to public health and risk. The analysis integrates a combination of methods similar to the previous two chapters, using keywords, multiword units, collocations, and dispersion of specific keywords. Major differences are prevalent in communication via Twitter, and the NHS accounts employ the most diverse messaging in terms of linguistic patterns and keywords.

The combination of methods in each chapter provides additional details and levels of analysis to evaluate each genre of language use. This produces further evidence for examining and understanding changes and similarities across each context. These corpus-based methods provide the tools for analyzing each of these key and important datasets.

Identifying and investigating the different ways that language patterns express views of health, public health concerns, and health-related technologies is of utmost importance. This work is important both for linguists and linguistic understandings of patterns and discourses in these specialized genres but also applicable to a wide audience of public health experts and to any individual in the wider public. As Berridge notes, "the public is confused about public health and uncertain what it means in practice" (2016:1).

This work exposes microlevel and macrolevel changes and trends by examining different types of corpora and sources. "Change over time is the key to understanding public health" (Berridge 2016: 4), and this is found in the linguistic patterns in each dataset. Analysis of the *Phil Trans* in the first English-language scientific journal reveals discourses and patterns surrounding the historical foundations of developing science and health discourses, as well as the foundations of understandings of vaccines, a fundamental problem as public debate continues concerning vaccines (British Academy 2020). This chapter illustrates the importance of information development and networks, in addition to the actual changes in linguistic patterns and keywords to the proliferation of health and scientific ideas. Key findings include changes in genre norms over time with the most changes in linguistic patterns culminating in the 1800s characterized by an increase in technical vocabulary, a greater focus on public health broadly, and descriptions of recognizable modern scientific processes. This chapter also illustrates the development of health-related ideas through keywords and collocations. Early Modern ideas about health were gradually incorporated and ultimately left behind as scientific knowledge developed and medicine was viewed as a practice. The consequences of science on health and medicine are demonstrated throughout the linguistic developments across this chapter.

Key findings from analyzing US and UK press on e-cigarettes present an interesting case of changes over time and press depictions of contested health technologies. The linguistic patterns across the ten years of corpus data display ongoing trends and differ between sources and locations. Differences include the US preoccupation with youth, business, and markets, along

with US-internal concerns. The UK press tends to focus on worldwide concerns in addition to issues within the UK, references people as users more generally, and is attentive to outside research related to e-cigarettes. One similarity, however, is the predominant lack of overtly negative or positive portrayals of e-cigarettes in US and UK press. Corpus-based analysis of media provides a key source of understanding helps with public perception and trends, as well as larger geopolitical and governmental differences.

Results from the Twitter COVID-19 governmental communication also underscores the importance of this research. Each governmental organization includes developmental changes in communication across 2020, but core differences remain between each group. The CDC initially focuses on information rather than offering advice, gradually providing more recommended actions later on in the pandemic. The WHO accounts concentrate overwhelmingly on different regions, cases, and COVID-19 itself throughout the year, while the NHS provides the most diverse messaging overall, focusing on individuals and groups of people, information, and recommended actions in a variety of keywords and linguistic patterns.

Each source of data, scientific articles, news press, and governmental tweets provide additional views into the core issue of public health communication. Tensions between scientific and public health authorities, governments, media, and the public are inherent threads throughout the analysis. "The narrative(s) of a crisis, as depicted by the media, can greatly influence how society understands and responds to a crisis" (Veil & Sellnow 2015: 472). This work is vital and highlights the necessity of continued research on related topics, with implications for supporting future public health endeavors through the thoughtful distribution and dissemination of public health information[1] (British Academy 2020).

[1] The British Academy recently reported about the importance of public support, dialogue, and accurate information with regard to public health and COVID-19, so that individuals can be fully equipped to make important health decisions. This only further underscores the relevance of this work.

CHAPTER 2

PUBLIC HEALTH COMMUNICATION

Communication is a key aspect of public health and interventions aimed at combatting health issues on an individual, communal, and global level. Language "not only reflects but also shapes the ways that people think and communicate about health" (Fox 1993, in Brookes et al. 2018: 99). Health communication is "essential to promoting and protecting the health of the public" (Bernhardt 2004) and also holds a central place in "risk and wellness" (Bernhardt 2004, Haider & Rogers 2005).

The field of public health communication is transdisciplinary and interdisciplinary. Bernhardt defines it as:

> the scientific development, strategic dissemination, and critical evaluation of relevant, accurate, accessible, and understandable health information communication to and from intended audiences to advance the health of the public (2004).

It is applied and centers on supporting positive health outcomes and behaviors (Haider & Rogers 2005) for communities rather than "deconstructing the underlying mechanisms of communication" (Bernhardt 2004). In terms of public health more broadly, a focus on health communication is only a recent trend (Bernhardt 2004). Many public health professionals view it as "more skill than science" with expectations that important health information would and could "speak for itself" (Bernhardt 2004). However, supporting any public health endeavor necessitates "both sound science and effective public health communication" (Bernhardt 2004).

There are many critical areas and key components of public health communication, practice, and methodologies. Public health communication plays a central role in the implementation of policy and community initiatives, in addition to aiding clinical and other

related health settings (Hornik 2002, Bernhardt 2004, Real & Buckner 2015, Bylund & Koenig 2015, Brookes et al. 2018). It incorporates diverse sources and practices aimed at individuals, communities, the public, experts, and politicians "with messages to impact quality of life" (Haider & Rogers 2005: xxvii).

Public health communication practice operates on a "continuum" in terms of goals and approaches, "from strictly individual change strategies to strictly social change strategies" (Maibach & Holtgrave 1995: 234). Regardless of the specific context or targeted audience, "effective health communication is critical for attainment of a broad range of health outcomes" (Merrick 2005: xxv). "Interaction" is a core aspect of public health communication, between health behaviors and choices, and the specific situations of communication (Haider & Rogers 2005). Often the most effective efforts in public health communication "focused on the problem of developing high quality messages reflecting particular evidence about the underpinnings of health behavior" (Hornik 2002: 13). This incorporates understandings of health behaviors and contexts. Targeted interventions are some examples of health communication aimed at impacting behavior to "address risk factors and ultimately reduce the burden of disease" (Haider & Rogers 2005: xxvii). Due to the broad nature of health communication, there are multifaceted associated practices including media campaigns, risk and crisis communication, community-focused interventions, and clinical practice (Maibach & Holtgrave 1995, Harrington 2015, Brookes et al. 2018).

Media campaigns are one key area of health communication, based in part on public health practices of "community advocacy, coalition building, and leadership development" (Maibach & Holtgrave 1995: 226). It refers to the employment "of mass media to advance a social or public policy initiative" (Maibach & Holtgrave 1995: 226). Media advocacy overlaps with related areas of public health communication, like marketing campaigns which are generally targeted at individuals (Maibach & Holtgrave 1995). Mass media provides a core context and

venue[2] for explaining "societal level causes" for public health issues and for spreading awareness to communities and individuals (Maibach & Holtgrave 1995).

Hornik points out that the efficacy of "public health education and communication on health outcomes" is not always clear-cut (2002: 1), noting that some community-focused trials have demonstrated little to no effects[3], while other studies[4] display "major change[s] in health behavior" (2002). Some examples of formal efforts utilizing mass media for public health communication include preventive efforts with AIDS (Hornik 2002: 8). Hornik compares U.S. efforts with other locations including Sweden and the Netherlands; despite the fact that findings from the U.S. communication efforts were "not as favorable as the Swiss or Netherlands data," a large effect on the AIDS epidemic persisted (Hornik 2002: 8). Additional examples of favorable outcomes have occurred in other situations, including "public education campaigns, national press, and media advocacy" for supporting and improving children's health in issues like Sudden Infant Death Syndrome (SIDS) and Reye's syndrome (Hornik 2002: 9).

Exposure also plays a key role in health communication (Hornik 2002: 13-14). Public health research suggests that utilizing "multiple communication channels" often and frequently while "encouraging natural social diffusion of messages" are important for enhancing exposure (Hornik 2002). In addition to careful messaging, "exposure strategies" should be based on goals and expected outcomes for enabling and promoting positive health behaviors and outcomes (Hornik 2002: 14). Public health communication campaigns include three models of behavioral

[2] Another related area of health communication incorporates the use of "interactive decision support systems (IDSSs)" (Maibach & Holtgrave 2005: 232). An interesting example of this is the CDC and other governmental unit's informational websites. they are also utilized by nonprofits to spread awareness and provide information in an easily accessible and rapid manner (Maibach &Holtgrave 1995: 233).
[3] Hornik argues that many of "these trials were not appropriate approaches to testing the effects of public communication; they were not able to provide very much increase in exposure to treatment when the treatment and control communities were compared" (2002: 10). Hornik argues further that more complex models of change are necessary for understanding the full effects of various health communication strategies on outcomes (2002: 11).
[4] These studies included both targeted programs and "normal media coverage of health issues" (Hornik 2002).

change, including those focused on individuals, social diffusion in public norms, and institutional diffusion (Hornik 2002: 14).

The invention of the internet heralded major changes in governmental strategies for communication regarding public health information (Yun et al. 2016: 68). These changes led to specific ways that entities could communicate and persuade the broader public by recommending and enacting "targeted action" and "manag[ing] public health crises more efficiently through direct public interaction" (Yun et al. 2016: 68). Social media enables this through specific "campaigns in order to meet their objectives of both informing and persuading a large swatch of a given population" (Yun et al. 2016: 67). Social media sites, including Twitter, have been used increasingly for "seeking and exchanging health information, contributing to a tremendous amount of health information available online" (Duggan et al. 2015, as quoted in Park et al. 2016: 188).

Risk communication is an additional core arena for public health communication (Maichbach & Holtgrave; Veil & Sellnow 2015) with far-reaching implications. Crisis and risk narratives are core aspects of media portrayals and public health communication and can shape public awareness and recognition of these narratives (Veil & Sellnow 2015). When different concerns are emphasized more distinctly, there are pitfalls. A distinct example of this is from the 1990s, when "the presumed benefits of a low-fat diet were emphasized by governments and public health professionals" (Kabat 2017: xv). This emphasis in communication is now traced to a variety of health concerns like "increasing rates of obesity" (Kabat 2017: xv). Furthermore, "the exaggeration and distortion of health risks can lead to the formulation of well-intended but wrongheaded policies that can actually do harm" (Kabat 2017: x). These distortions can have further implications for audiences and individuals who "may become desensitized to health messages" as the result of an overload of distribution of "the latest imaginary threat" (Kabat 2017).

Risk communication with health issues is also vital in crisis situations (Veil & Sellnow 2015) like the COVID-19 pandemic, leading to issues in the description and frequency of scientific findings and updates (Collins & Nerlich 2017: 291-2, Larson 2020: xxxvii). Large amounts of data, and scientific knowledge are distilled and even further complicated by public perceptions of risk and uncertainty, often embroiled in the politicization of health risks and decisions (Collins & Nerlich 2017: 291-2, Larson 2020: xxxvii). Health messaging is impacted by subjective judgement, with implications for disentangling evidence and interpretation (Kabat 2017: xvii).

Another integral aspect of health communication research involves examining "what health and illness mean to local communities", thereby considering relevant cultural influences on these areas (Ho 2015: 233). Health, perceptions of health, and "definitions of health" are not static (Berridge 2006, 2016); instead, they are "profoundly affected by the social, political, environmental, and behavioral factors with which people live" (Hornik 2002). Therefore different strategies and approaches are followed depending on the "level" and audience awareness goals (Hornik 2002). This communication is also generally utilized with other policy and advocacy actions to support goals and outcomes.

When public health concerns arise, it is imperative that the units involved (whether nonprofits or governmental agencies) "seek the involvement of the affected or concerned communities" (Saunders et al. 2006: 137). Although many programs often use unilateral communication in different initiatives and media campaigns, others utilize two-way and interactive communication with consumers and audiences. This coordination "between sources and receivers to ensure that messages are accessed and understood, communities are involved and invested" and programs can be adjusted where needed (Bernhardt 2004). The unilateral contexts[5] of health communication from practitioners to individuals brings up interesting and valid

[5] Bernhardt notes that this unilateral nature can be viewed as "paternalistic" (2004).

concerns about the efficacy of such communication. Regardless, when thoughtfully "implemented" and deployed,

> public health communication programs have the capacity to elicit change among individuals and populations by raising awareness, increasing knowledge, shaping attitudes, and changing behaviors. Although communication initiatives often target for change those behaviors that contribute directly to morbidity and mortality, public health communication also targets social, physical, and environmental changes that can influence health outcomes (Bernhardt 2004).

The central tenet of public health communication is to present health information in a clear and persuasive manner for "individuals and society" (Regidor et al. 2007: 93). This is highly involved with important considerations for contexts and explanation of the science behind health decisions (Collins & Nerlich 2017: 291-2, Larson 2020: xxxvii). Media influence is also persistent, and the "boundary between what constitutes a health risk can quickly become blurred by speculation in the media" (Tang & Rundblad 2017: 667). Situations of intense pressure and uncertainty like the COVID-19 pandemic have major impacts on public health communication and interpretation (Larson 2020: 47, Collins & Nerlich 2017: 291-2). Regardless of the specific context, methodology, or targeted audience, there is "no strict blueprint for effective clinical" or health communication[6] (Brookes et al. 2018: 111).

[6] Clinical practice in health contexts and the "methodological principles of corpus linguistics" are well-suited in combination to provide "pedagogical interventions...[that fit] medicine's ethos of 'evidence-based' communication/practice" (Brown, Crawford, & Carter 2006 as quoted in Brookes et al. 2018: 109).

CHAPTER 3

CORPUS LINGUISTICS

1. Theoretical and Applied Background

This work analyzes trends in English language variation in scientific and health communication across distinctive datasets from Early Modern English in the first English scientific journal (*The Philosophical Transactions of the Royal Society of London,* Fischer et al. 2020) and present-day genres of news press and social media (Kuiper forthcoming). Although this endeavor is primarily situated in corpus linguistics, an approach to linguistic inquiry that utilizes "machine-readable collection[s] of texts" (Kretzschmar 2018, Gries 2017: 7), otherwise known as corpora, other interdisciplinary work is included from digital humanities and computational studies. Corpus linguistics involves many different areas of research including semantics, lexicography, and language variation, with far-reaching applications to science and health messaging and change in communication over time (Kretzschmar et al. 2004). Corpus linguistics has evolved and expanded since its initial role in linguistics; McEnery et al. (as quoted in Baker 2011) note that it is not an "independent branch of linguistics in the same way as phonetics, syntax, semantics, or pragmatics" but closely relates to other subfields and interdisciplinary work. Corpus linguistics has proven advantageous for a wide range of fields and interdisciplinary efforts in linguistics and beyond, such as applications to science and health communication and differences in linguistic patterns over time[7]. Wynne argues that "corpus linguistics offers some of the most powerful new procedures for the analysis of language, and the impact of this dynamic and expanding sub-discipline is making itself felt in many areas of language study" (Wynne 2004: 3).

There are several key studies underpinning this endeavor, most notably research on corpus linguistics and complex systems[8]. Kretzschmar notes that

> language in use…is first and foremost not a logical system but a complex system [and] emergence in English is not once-and-done; it continues in every place where the language is spoken or written, in every locality and in every kind of conversation or text (2018: 2).

Emergence is what underlies linguistic variation and communication in any situation, including the text types analyzed in this work (Kretzschmar 2015: 212-3, 2018: 3-4). Complexity science makes an advantageous fit with corpus linguistics because it provides scientific and rigorous ways of dealing with and analyzing linguistic data. Studying language as a complex system also allows the researcher to consider additional influences like culture and institutions on language change through interaction and emergence while underscoring that "language change does not always proceed at the same rate" (Kretzschmar 2015: 212-3, 2018: 32). These theoretical foundations encompassed in emergence and interaction inform all aspects of this work, from decisions in the corpus-building process to each section of the analysis and discussion.

Additional work emphasizes cultural influences as fundamental aspects of linguistic change and knowledge (Firth 1935, 1967, Stubbs 2001, 2002, Baker 2006, 2011, McEnery & Baker 2017). Stubbs (2001:10) reports that linguistic knowledge is not just about single tokens but instead includes "their predictable combinations" and further the "cultural knowledge which these combinations often encapsulate". Another way to describe this is through Firth's famous and oft-used quote "you shall know a word by the company it keeps" (Firth 1957: 11). Firth prioritized the study of "language in use" (Kretzschmar 2009: 11), arguing that "the key to a better understanding of what language really is and how it works" is present in actual linguistic situations of use, including speech (Firth 1935: 66, 70-71, as quoted in Kretzschmar 2009:11).

[8] Specific research methods related to this are discussed in more detail in section 3 of this chapter.

This also means prioritizing the "situational context" of the interaction (Kretzschmar 2009: 148) when analyzing the linguistic situation. Stubbs describes this by observing that linguistic patterns vary depending on the context (Stubbs 2001: 11-14), so "different text types have different patterns of expectation" (Stubbs 2001:23). These patterns play out on different levels, including the phrasal level (Stubbs 2002: 3), grammatical level, social context, and larger cultural discourses. As Kretzschmar points out, corpus analysis is a "prominent expression of Firth's ideas" and corpora are electronic preservations of "authentic examples of language in use" and make it possible to interrogate different layers of meaning and language patterns (2009: 151-2).

Further, language variation and lexical change in particular presents interesting opportunities for study as it

> "has the potential to tell us much about societal change. Language does not develop in isolation but has a dialectical relationship with culture, both reflecting and spurring on changes in every day life" (Baker 2011).

Stubbs notes a particular example of context and culture in the context of the word *bank.* He highlights speakers' "unconscious knowledge about expectation of language patterns" (2002: 20). The example of the word *bank* in "isolation and in different contexts" underscores the semantic idea of ambiguity and the role of the phrase (2002: 20-21). At the phrasal level, the different meanings associated with *bank* are inferred by the audience: "the supermarket is opposite the bank" and "river bank" (2002: 14). This is a small example of the influences of linguistic patterns and contexts in addition to the larger cultural context influencing meaning and linguistic expectations.

Other key scholars have worked on text analysis methods in additional contexts and genres, including literature. Digital humanist and literary scholar Burrows argues that "computer-assisted textual analysis can be of value in many different sorts of literary inquiry" (2004: 323-4), and Jockers notes that the advent of corpora and "large-scale book digitization projects" have changed the "nature of evidence" available to scholars (2013: 6). Jockers points out that this new

evidence "invite[s] and even demand[s] a new type of evidence gathering and meaning-making" (2013: 6). Burrows, Jockers, and others are forerunners in this area primarily in digital humanities and literary studies. Burrows utilizes various methods of classification and grouping texts through use of cluster analysis, for example (2004). His work analyzing 17th and 18th century English poetry displays ongoing "differentiation" between early and later works, with specific authors like Samuel Butler standing out due to their "idiosyncratic nature" highlighted through cluster analysis (2004: 326-7). Burrows discusses implications for what decisions are factored into the analysis, like whether to include lexical or function words (2004: 328). Jockers similarly applied methods including cluster analysis (2013: 90-104), frequency analysis of lexical and other features, and network analysis to literary genres (2013).

Many digital humanities scholars trace the first large, computational text analysis study to Father Robert Busa's concordancing of the works of Thomas Aquinas (Jockers 2013: 3, 11-13). Another notable dataset is the Brown corpus, created by Henry Francis and Nelson Kucera; it is one of the first examples of a carefully sampled and compiled corpus (1961). It includes representative excerpts of American English across fifteen different genres and is a well-known and recognized general corpus (Kretzschmar 2018: 218). Additional general corpora of note are the British National Corpus (BNC), a 100 million word general corpus of British English (2007) and the Corpus of Contemporary American English (COCA) which contains over 1 billion words of general American English (Davies 2008). Each provides distinctive samples of British and American English, respectively, across different categories and text types. These corpora were carefully compiled and created with additional metadata included, such as text types and linguistic annotations like part of speech tags. COHA, the Corpus of Historical American English is another notable corpus and was developed by Davies (2010). It contains over 400 million tokens from genres over time from 1810 to 2009. "Different corpora represent different populations of texts" (Kretzschmar 2018: 218) so that researchers may draw conclusions based on the specific sample and population. Specialized corpora are designated as such based on their

compilation and sampling[9] criteria and are also useful for an abundance of notable research questions.

A central area of language change and corpus research involves the analysis of health communication and health-related data. Applications of corpus linguistics to these arenas have been widely recognized (Atkins & Harvey 2010, Baker et al. 2019), though only recently have been applied. Indeed, the most formative applied study underpinning this project is Kretzschmar et al. (2004)'s analysis of communication and linguistic variation in a corpus of US Tobacco Industry documents. This study is particularly effective for many reasons, and it is one of the first large scale corpus analyses of public data. Through keyword analysis and principled sampling, unique characteristics of linguistic and rhetorical strategies of company-internal messaging are found, in addition to the effectiveness of principled sampling methods, and the "extreme paucity of external audience documents" (2004: 54). This study also succeeds in demonstrating the merit of corpus-based research on public health and scientific communication and useful strategies for corpus design and representative sampling procedures (Kretzschmar et al. 2004).

Other corpus-based studies on health communication include work on specific contexts and genres, including patient-doctor and patient-healthcare professional interactions, such as work by Atkins et al. on the Nottingham Health Communication Corpus, and Baker et al. 2019 on the National Health Service (NHS) Communication corpus. Atkins and Harvey note that corpus linguistics applies particularly well to the study of health communication and health-related data (2010: 606), including studies on communication in institutional health-care settings, doctor-patient interactions (Patra et al. 2015) and medical texts and discourses, such as case-histories and note-taking (Francis & Kramer-Dahl 2004, as cited in Atkins & Harvey 2010: 606). These studies provide thoughtful insight into the communicative efficacy of the NHS. One of the first studies on health care communication focused on NHS professional-patient interactions with cancer

[9] See pages 63 and 92-93 for more discussion on this.

patients, finding specific linguistic features that professionals used when discussing health with their patients during consultations (Thomas & Wilson 1996). Atkins and Harvey's work on the Nottingham Health Communication corpus focuses on a wide range of health professional interactions with the public, including pharmacists and nurses. Methods for analyzing the latter corpus include keyword analysis with qualitative discussions, and findings involve evidence of health professionals' employment of vague language in practice (Atkins, Adolphs, & Harvey forthcoming, Atkins & Harvey 2010, Adolphs 2002, Adolphs 2004).

More recently, Baker et al. argue that integration of concordance lines is especially beneficial for understanding the specific contexts and linguistic patterns in patient and practitioner interactions and feedback, emphasizing detailed evidence of positive feedback overall in NHS patient practitioner interactions (2019: 215, 223-4). In comparing specific roles of NHS workers and feedback, they find that "surgeons and dentists tended to receive extremely positive feedback," while nurses and receptionists received less feedback overall (2019: 216). Out of all NHS workers, receptionists were the most likely to be viewed as "NHS villains" (2019: 216). Baker et al. also note that stylistic changes in patient feedback offer greater insight into "more subtle features linked to patient motivations, expectations, fears or self-constructions" (2019: 214)**.** Semino et al. have considered the use of metaphors in framing patient care, the efficacy of pain consultations, and other situations in a variety of healthcare interactions (2015, 2017, 2018, 2019). Adolphs et al. note that corpus linguistics in particular provides a "nuanced explication of communication dynamics or 'linguistic signatures' directly associated with a variety of health care interventions or inputs" (nd). Another relevant study focused on the SARS crisis in 2003 and analyzed semantic prosody with respect to collocation, colligation, and semantic preference in the Hong Kong Corpus of Spoken English and found that the patterns of co-selection provide further examples of textual and intertextual coherence (Cheng 2009). In addition to communication in specific healthcare contexts, corpus studies have also focused on variation and genres that discuss health topics in different registers.

The connections between linguistic patterns and societal changes have been studied extensively in work related to genres of language use, including media corpora and press accounts of health-related representations, with noteworthy findings regarding linguistic characteristics across text types and genres (Biber et al. 1999, Biber et al. 2004, McEnery et al. 2006). This research underscores the useful insights that corpus linguistics offers into repeated patterns of language and the representations of groups, ideas, and concerns in the press (Mautner 2008, Baker et al. 2013, Gupta 2015: 5).

A central area of cultural influence includes changes in technology and information networks. The development of the printing press and trade facilitated and reinforced long-term networks of distribution of not only physical goods but of information (Matson 2019: x). Eliot and Rose argue that not only do "books make history" but they are also "made by history…shaped by economic, political, social, and cultural forces" (2007: 1). They are one example of the way that technologies become tools to communicate and share ideas and narratives, to "exercise power and distribute wealth" (Eliot and Rose 2007:1, 9; Ho 2015). With the integration of the printing press into society, scientific work and practice became increasingly motivated by both intellectual and "commercial motives", impacting scholars and intended audiences (Taavitsainen 2017, Culpepper & Kytö 2010: 24). The printing press helped facilitate the growth of popular journalism and wide readership, resulting in substantial increases in English prose and the development of different genres (Culpepper & Kytö 2010: 39). Tufekci notes that technological interactions are not fixed, and "everything evolves as people invent, innovate, and appropriate technologies for their purposes" (2017: 263). Institutions and larger cultural discourses were and are still informed by technologies, especially given that "technologies alter our ability to preserve and circulate ideas and stories, the ways in which we connect and converse*"* (2017**:** 7). This also impacts how people share and understand new information and evidence (2017: 7); technologies facilitate the growth of communication in new genres and methods, providing important data sources (Tufekci 2017: 7). Tufekci discusses

several parallels in early technologies such as the printing press with "digital connectivity" technologies, noting that issues of power and "misinformation have always been part of the public sphere" (266). The printing press was utilized by the Catholic Church and to challenge the authority of the Catholic Church and received knowledge, much like internet and media genres of language use today are used by various groups to challenge power hierarchies and to promote them (2017: 270).

Each genre studied in this work is a result of the interaction between cultural institutions, scientific developments, and technologies. Different language situations, genres, and text types have expected conventions in terms of internal and external characteristics and diverse expectations for audiences (Taavitsainen 2017: 253-4). Genres are often studied in terms of "text external features," and text types are distinctive through internal linguistic evidence (Taavitsainen 2017: 256). Taavitsainen argues that genres should also be considered in terms of their emergent history, pointing out that in the context of the evolution of English, "the first attempts to transfer learned genres from classical models to ME were made in the fifteenth century" (2017: 257).

Author intentions are also part of the "meaning-making mechanisms inherent in genres" (Taavitsainen 2017: 253, Baker 2011: 13) and variation in discourse and microlevel patterns. In some cases, the distinctions between audience and author and information and opinion are muddled (Diamond 2012: x). Diamond emphasizes this further by noting that the advent of digital communication technology has dramatically impacted communication and audiences:

> Few, if any, developments in recent decades have more profoundly transformed politics and civil society than the emergence of digital information and communication technologies (ICTs)…These electronic tools have provided new, breathtakingly dynamic, and radically decentralized means for people and organizations to communicate and cooperate with one another for political and civic ends (2012: iv).

A key area of digital communication is the area of risk and scientific communication. Tang and Rundblad note that language structures play an integral "role in the portrayal of risk and science"

(2017). Scientific and health communication are notable data sources in terms of audience impact; much work has been done on press power to "shape widely shared constructions of reality" and influence readers regardless of type of media (Jaspal & Nerlich 2017, Mautner 2008: 3). Each genre involves varying sets of expectations and goals, representations, and characteristics (Baker 2013: 3).

Another notable text type in this work is social media language, and Baker argues that internet or "web language" is a variety or genre of language in and of itself (2011: 13). Social media is a facet of internet language, and a key area of research on science and health communication. "Social media have become important information hubs, where individuals and organizations create and disseminate real-time content beyond their personal social networks and physical location" (Gui et al. 2018: 820). Bednarek points out that sharing information through social media "is an increasingly important consideration in journalism practice and production" (2016: 227-8). They are also exceedingly powerful and influential during times of crisis worldwide (Diamond 2012: x, Gui et al. 2018: 820); however, utilizing social media "for any public health crisis requires complete consideration and dedication" (Juyal et al. 2020).

Different language situations and methods have made great strides in studies of scientific and medical communication. These studies demonstrate the wide-ranging applications and findings using corpus-based, quantitative, and qualitative means. Further work remains regarding the impact of linguistic patterns on science and health communication, including the interplay of interactions and emergence in the language context.

2. Considerations for Corpus Creation

This section describes the methods and considerations for corpus compilation and creation, data gathering, and analyses employed in chapters 4-6. A variety of methods and tools are implemented in data preparation and management. One reason why corpus-based linguistics offers such fruitful possibilities is because of the considerations involved in creating corpora.

There are many relevant factors when it comes to obtaining data for corpus compilation. Thoughtful design is a crucial step of the process (Kretzschmar et al. 2004, Kretzschmar 2018, Egbert et al. 2020). Throughout the workflow, programming and software logistics, careful notation of versioning and metadata, and linguistic considerations of representativeness, sampling, size, and balance are integral. Sinclair notes that a corpus ought to be created with selection based on "external criteria" or based on the "communicative function of a text" (2004). A corpus is in itself "a remarkable thing, not so much because it is a collection of language text, but because of the properties that it acquires if it is well-designed and carefully constructed" (Clear 1992, as quoted in Sinclair 2004). These properties are key aspects of the differences in corpora and other examples of Big Data (Jockers 2013).

Representation is one significant facet of the corpus creation process. Representation is based on whether a corpus delivers a useful sample of the population or snapshot of language it is meant to describe (Brezina 2018: 15). Sinclair notes that no corpus can totally or completely represent a language, language variety or genre, but it must be as representative as possible within reason for the context it is meant to exemplify, as "no limits can be placed on a natural language" (Sinclair 2004). Egbert describes this well, noting that the main aim in corpus linguistics

> is to carry out research on a corpus of texts that is as representative as possible of a target population of interest. Corpus linguists are interested in how language is actually used in a register, dialect, or entire language (Egbert et al. 2020).

An exemplary corpus in terms of representation, balance, and sampling is the first general corpus of American English, the Brown corpus created by Kucera et al. (1961). Brown was created to represent a snapshot of American English in the 1960s and contains a selected sampling of different language contexts and genres, including literary and non-fiction text types. Brown is a general corpus representing American English in this selected sampling frame which includes a total of five hundred 2,000 token text samples (Francis & Kûcera 1979, as quoted in Brezina 2018: 16). The corpora analyzed in chapters 3-6 are specialized corpora which are

classified according to various terms, including genre, author, time frame, and others. Both general and specialized corpora are advantageous for various types of analyses. General corpora differ from specialized corpora in the areas of representation and selection for inclusion in the dataset (Kretzschmar 2018: 218-9). Sinclair notes that the "structural criteria" selected for corpus creation must be applied throughout the process, providing a "framework for the principal corpus components" (2004).

Balance is another aspect of corpus creation and refers to considerations involved in criteria for inclusion in a corpus (Wynne 2004, Sinclair 2004)**.** Although the notion of balance in a corpus is more clearly associated with creating general corpora (Kretzschmar 2018: 219), balance applies to any effort of corpus creation because it is "essential that the structural criteria" of the corpus be carefully considered in the building process for further analyses and work (Sinclair 2004). For any dataset to be considered a balanced corpus, "the proportions of different kinds of texts it contains should correspond with informed and intuitive judgements" (Sinclair 2004). Any conclusions based on a corpus relate to the overall balance and representativeness of the corpus itself. Regardless of the researcher's decisions in reaching the final constructed corpus, the implementation of corpus design and finalized construction "should be documented fully", including reasons for its composition (Sinclair 2004).

Random sampling is one way to ensure that bias is removed from the corpus compilation process, but it is not always feasible to achieve this, depending on the dataset (Brezina 2018: 15). In certain instances, the corpus can include the full population of interest, as in Baker et al. 2013's study of press, Partington's study of antisemitism in the British press (2012), and the Royal Society Corpus (RSC) utilized in Chapter 6 (Fischer et al. 2020). Kretzschmar et al. 2004 refer to the use of principled sampling, which enables valuable analyses to follow**. "**Corpus designers often start with a set of categories, within which they aim to collect an unbiased sample" (Brezina 2018: 15-6).

Dataset size is another related factor in corpus building which crucially depends on many issues, including the types of queries and the methodology involved in studying the data (Sinclair 2004). “There is no maximum size” of corpus, but instead the size of the corpus, in addition to all decisions related to sampling, balance, and representation must be supportable and well-justified. Sinclair notes that collocational patterns are one way to decide on the size of the corpus: “The density of the patterns of collocation is one of the determinants of the optional size of the corpus. Other factors include the range of ambiguity of a word chosen, and sometimes its distribution among the corpus components” (Sinclair 2004: 18). Regardless of the final choices made by the corpus compiler, “the more [corpus] data you can gather, the clearer and more accurate will be the picture that you get of the language” (Sinclair 2004). The corpora used in this work are a variety of sizes, with the RSC being the largest corpus at 78 million tokens (Fischer et al. 2020).

This work further entails the creation of multiple specialized corpora. Compiling specialized corpora with data involves specific selection and inclusion criteria, addressing each consideration discussed above, in addition to careful metadata gathering and versioning practices. The selection criteria for the corpora in this work are discussed more thoroughly in each chapter; however a brief summary is also provided below.

3. Corpus Compilation

Key considerations for the creation of corpora for the first two analysis chapters include metadata storage and processing, along with annotation of the data. Setting up and organizing both corpora involves similar processes and scripting, preserving each version of the corpus after adding additional data. Kretzschmar et al. note the importance of data preservation and principled sampling to ensure that the final corpus versioning will enable productive results (2004). Corpora are often compiled with metadata, grammatical, and linguistic information. In the CWB-encoded format, these annotations are referred to as structural (s)-attributes (metadata) and positional (p)-attributes (linguistic information) (Evert & Hardie 2011, Evert et al. 2021). All corpora in this work are encoded in CWB-format (Evert & Hardie 2011, Evert et al. 2021) and utilize CQP and

R for analyses. S-attributes are appended to each word, and these vary per corpus discussed above. Some examples of this information include date and year of publication, author, source, and geographical location. P-attributes include lemma and part of speech (POS) tags. The part of speech tags in the creation of the Twitter corpora of CDC, NHS, and WHO accounts are from the cleanNLP package in R (Arnold 2020), which employs the CoreNLP package from the Stanford Natural Language Processing group (Arnold 2020, Manning et al. 2014, Toutanova et al. 2003) and the Penn Treebank Tagset (Marcus et al. 1993).

Corpus preparation and compilation is accomplished with a variety of computer programs: Python (Version 3.7.4.), Unix, CWB (Corpus Workbench), R Studio (R Core Team 2019), and R packages like tidyverse (Wickam et al. 2019), textmining (Feinerer 2020) and tidytext (Silge & Robinson 2021) and others for data preprocessing. Data processing and formatting is another core step in corpus creation and requires a pipeline of programming and manual steps to prepare and achieve the final version of each corpus.

4. Corpus-based Analysis Methods

"Meaning is found in discourse, in the form of textual evidence" (Teubert 2019: 140). Different types of textual evidence are used as critical methods from corpus linguistics, including frequency analysis, analysis of collocations and concordance lines (keywords-in-context KWIC). These methods are distinguished for identifying linguistic and rhetorical patterns in corpus data and are inextricably connected with cultural issues, discourses, and ideologies (Hunston 2010: 162-166, Fairclough 2003, Kretzschmar 2018: 218, Wiegand & Mahlberg 2019: 5). Stubbs writes that "meaning is use" (2001: 13) and employing these methods allows researchers to record and understand the way tokens and linguistic patterns "acquire or change meaning according to the social and linguistic contexts in which they are used" (Stubbs 2001: 13). Corpus analysis provides insight and evidence into "language in use" and "social interactions and behaviors" in different contexts (Weigand & Mahlberg 2019: 1). "Situational and cultural parameters involved in which a text is produced are not fully reproducible from a text, but are reflected in its lexico-

grammatical patterns" (Wiegand & Mahlberg 2019: 4); the linguistic patterns evidence the larger cultural influences and discourses in any given text or corpus.

Analysis of frequencies includes single token units, n-grams, and collocations. N-grams are repeated token patterns of any length and provide insight into how different genres and discourses use language structures (Scott & Tribble 2006: 132, Römer 2010). Collocations are co-occurring words and involve a span of up to 5 words on either side of the node word of interest. Along with n-grams, collocations have been utilized in traditional corpus studies as a variation of types of frequency analyses (Scott & Tribble 2006, Scott 2010: 148, Römer 2010). Although frequency analysis is featured in each chapter, ngrams are analyzed in chapters 4-5. N-grams are similar to collocations in that they are sequences of tokens following each other, a variation of frequency analysis (Scott & Tribble 2006: 132, Römer 2010, Scott 2010). Different variants of ngrams are analyzed as bigrams (two tokens), trigrams (three words), and sometimes four-word tokens. A characteristic of the output of each of these types of analyses is that the frequency profile will always follow the asymptotic distribution and Zipf's law (Sinclair 2004, Kretzschmar 2009, 2015, 2018, Brezina 2018). This applies to frequencies of function words, content words, collocations, or n-grams.

A distinctive method in corpus linguistics is the analysis of collocations, which are words that occur together. There are different metrics and properties involved in examining collocates. Some are preferable for measuring the strength of a collocation, while other metrics are preferable for identifying the most typical associated collocations. Collocations are considered in terms of frequency and different spans surrounding the node word. Association metrics between collocations utilize different statistics to understand the strength of these associations (Brezina 2015: 67-69, Gablasova et al. 2017, Gries 2013) via different methods (Dunning 1993, Gablasova et al. 2017, Gries 2013). These methods include commonly used association measures like mutual information scores (Gabrielatos & Baker 2008, Gablasova et al. 2017) and log-likelihood (Dunning 1993, Gabrielatos & Baker 2008, Blätte & Leonhardt 2019). Log-likelihood and

frequency ratios are used in the queries and functions across this work. Log-likelihood is distinguished for providing the most frequency associated collocates (Baker 2014: 134-5), while mutual information scores are assigned based on "strength or salience of collocation" (Baker 2014: 134-5, Gablasova et al. 2017).

Crucially, collocations reflect underlying discourses and ideologies associated with the corpus (Fairclough 2003); they also expose the patterns of meaning in use of a particular corpus (Kretzschmar 2018: 220-1). The frequency profiles of collocations highlight the semantic prosody associated with the node word (Kretzschmar 2018: 221). McEnery and Baker also describe further ways that collocations form or are distributed across a corpus (2017), in terms of "consistency", "initiation", "termination", and "transience" (2017: 27-8). Consistent collocates are those that occur with the node word for a specified consistent amount of time; for McEnery and Baker, this was for at least seven of the ten decades being studied (2017: 26). Initiation refers to the idea of collocates that newly enter the corpus data and remain consistently present for a specified amount of time across the corpus (McEnery & Baker 2017: 26). Transient collocates are those "associated with a short period" of time before leaving the corpus data (McEnery and Baker 2017: 28). "Connectivity" is another property of collocations examined (Brezina 2018: 141).

Dispersion is another method that is used in different situations to analyze the change in use over time of certain tokens, collocations, and ngrams. Dispersion measures the distribution of a particular linguistic item in a corpus (Brezina 2018: 47) and is useful for analyzing collocations, including each type of collocation, as described above from McEnery and Baker (2017). There are different advantages for each method. Dispersion is useful for considering if particular tokens show up more frequently in certain contexts. There are no single measures for dispersion, but often ratios are utilized to compare across different sizes of corpora or subcorpora.

Analysis of keywords provides researchers with a particular viewpoint of the data, allowing for more textured and nuanced considerations of the data. McEnery and Baker describe this as providing information on the "'aboutness' of a large text collection' (2017: 24-5).

Keyword analysis is utilized by relevant work; perhaps described best by Kretzschmar et al. 2004 and featured in other notable work, including Baker et al. 2019, McEnery & Baker 2017, Gabrielatos & Baker 2008, Hannaford 2018, 2021. Keywords are tokens in a corpus that are "statistically significantly more common in one corpus than another" (Kretzschmar 2018: 219). Keyness is the statistic that is most often utilized, measuring "whether a word has a higher or lower frequency than expected in a subcorpus" in comparison with a reference corpus (Kretzschmar et al. 2004: 40). If a significant positive keyness occurs in any given token, then that token is statistically significantly more frequent in the subcorpus, while a negative score "indicates a much lower frequency of use for a particular word" (Kretzschmar et al. 2004: 41). These methods are highlighted in the RSC chapter; R packages enabling keyword analysis include the polmineR package (Blätte & Leonhardt 2019), quanteda (Benoit et al. 2021), and tidyverse (Wickam et al. 2019). Taavitsainen points out that this method in particular (with ngrams) "can tell a great deal about meanings" and discourses (2017: 255).

The last method utilized is the analysis of concordance lines, or keywords in context. This was one of the most notable core methods at the advent of corpus-based analysis, providing researchers with a novel way of viewing and analyzing linguistic data. Concordance lines feature different settings that allow researchers to view multiple contexts of a given token; in this work, the KWIC function from polmineR (Blätte & Leonhardt 2019) is implemented.

For each of the discussed methods, the required functions are executed via R packages discussed above. Generally, ggplot and tidyverse (Wickam et al. 2019) are the primary packages used for data processing and for data visualization, while polmineR (Blätte & Leonhardt 2019), tidytext (Silge & Robinson 2021), and quanteda (Benoit et al. 2021) are the main packages implemented for corpus-based methods like analysis of collocations, ngrams, and keywords.

5. Conclusion

Integrating a combination of methods is most often preferable to avoid researcher bias (Baker 2014, Brezina 2018) and to provide a more fine-grained viewpoint of the corpora being

analyzed. Taavitsainen notes that the implementation of multiple methods also provides advantages for revealing distinctive linguistic features in the data like illuminating "aspects of genre dynamics" (2017: 255). Baker argues that it is important for corpus linguists to utilize many techniques to avoid research errors and bias (2014: 200). Another key consideration related to this is the role of the researcher in interpreting corpus-based techniques and methods. Teubert emphasizes that although the array of methods involved in corpus-based linguistics are "irreplaceable" for finding patterns, "what it [corpus-based methodology] does not do is to venture an interpretation of its findings. Interpretation is what only people can do" (Teubert 2019: 141-142). This underscores the integral role of the researcher in the corpus-based process.

In the chapters that follow, a variety of each of these methods is implemented on different datasets across genres of science and health communication in English over time. R and each of the functions utilized to find frequent patterns, keywords, and phrases emphasize the changes and developments in communication and discourses in key areas of messaging. These patterns are integral, as they enable people to "make sense of the world" (Teubert 2019: 138).

CHAPTER 4

PROMOTION AND PRESERVATION OF PUBLIC HEALTH: TRENDS IN HEALTH AND SCIENCE COMMUNICATION IN THE ROYAL SOCIETY CORPUS

1. Introduction

Efforts in public health have been traced back to scientific advances from the beginning of the Royal Society (RS), and there are many distinguished members of the RS that made significant contributions in these efforts. This chapter examines discourses of health and medicine in the first English scientific periodical, the *Philosophical Transactions of the Royal Society of London* (*Phil Trans*), founded by Henry Oldenburg (Gross et al. 2002, Kermes et al. 2016, Fischer et al. 2020). The Royal Society Corpus (RSC) is a specialized and representative "diachronic corpus of scientific English" (Fischer et al. 2020: 794) of the *Phil Trans* from its beginning in 1665 to 1920 (Fischer et al. 2020); it is approximately 78 million words and has been encoded for text types, year of publication, and has been tokenized and linguistically annotated for lemma and part of speech using the TreeTagger (Schmid 1995). From its beginnings, the *Phil Trans* provided a new means for "disseminating the scientific information that provided momentum to the scientific movement that still continues today," with emergent and distinctive genre characteristics (Kronick 1990, Atkinson 1999: xxii, Gross et al. 2002: viii, Kermes et al. 2016, Biber & Conrad 2019: 222, Biber & Conrad 2019: 236-7).

The time span of the *Phil Trans* covers wide-ranging societal, scientific, and technological developments. At the beginning of the Journal, "the Aristotelian and Ptolemaic vision of the universe" was widely accepted (Tinniswood 2019: 11); even the most educated individuals continued to subscribe to the humoral theory of disease from the Middle Ages (Tinniswood 2019; Wootton 2015), and ancient theories of natural philosophy were taught extensively in places of

learning across the Western World (Tinniswood 2019: 12). The *Phil Trans* is currently "the oldest continuously-published science journal in the world" and established "scientific priority and peer review" (RS 2020, Biber & Conrad 2019: 236-7). The Royal Society, in addition to sharing new findings through the *Phil Trans*, "established a public forum for the sharing of ideas, and in so doing, they created an international culture of knowledge exchange" (Tinniswood 2019: 68). Today, the Journal includes topics of biophysics, evolution, epidemiology, engineering, and many others.

In the 1650s, authors of the *Phil Trans* describe *fever* as "Nature's Engine". This simple phrase demonstrates key departures from early natural philosophy to a greater understanding of processes in living organisms. It also represents a crossover between physical and life sciences and is a notable example of the interaction between cultural norms, scientific developments, and medical practice. As shown in previous studies (Kuiper forthcoming, Atkinson 1999, Biber and Gray 2010, Gotti 2011, McEnery & Baker 2017), fundamental changes in science align with linguistic changes in communication. This chapter focuses on the consideration of specific tokens *health, disease(s),* and *inoculate|inoculation* using collocational and concordance analysis to understand notable discourse trends. Each section of analysis is informed by and reflective of previous findings, including increased technical vocabulary over time, reduced use of first person, and others[10].

2. Related Literature

English gradually became the Lingua Franca for scientific communication (Gläser 1995: 188-9, Oliver del Olmo 2014, Kamadjeu 2019, Taavitsainen & Pahta 2011: xvii). The majority of learned texts from the Middle Ages to the early eighteenth century were written in Latin; however, if an author selected English as the language of publication, it was a loaded decision, "at a time when English was competing with Latin for public prestige" (Gläser 1995: 188).

[10] See Related Literature in Section 2, especially work by Gläser (1995) and Gotti (2011).

Indeed, Francis Bacon wrote *Of the Proficience and Advancement of Learning* in English in 1605, and then authored two other books in Latin. Sir Isaac Newton first published in Latin and then proceeded to write *Opticks* in 1704 in English (Gläser 1995: 189-90). These decisions, combined with the development of the printing press, the gradual development of a literate society and political and socio-cultural influences led to the proliferation of scientific publishing in English, which remains at present the predominant language of science (Götze 1997: 53, Kaplan 2001; Martel 2001: 27-30, Hyland, 2002 as quoted in Oliver del Olmo, Pahta and Taavitsainen 2011: xvii).

Gläser discusses the transformations in scientific communication from largely Latin-dominated to English, noting that the scholar's choice of writing in English was deliberate, "at a time when English was competing with Latin for public prestige" (1995: 188). Many prominent philosophers, statesmen, and others wrote English equivalents of their Latin work, including Thomas More, Francis Bacon, William Harvey, Robert Boyle, Robert Hooke, and Sir Isaac Newton (188-189), showing an increasing "preference for English in the world of science" (1995: 189). Gläser also includes an overview of linguistic and rhetorical characteristics of scientific writing in the 17th century, including the use of first person singular and the introduction and utilization of technical vocabulary (1995: 192-197). This is also found in studies of specific genres of scientific prose, like medical prose (Gotti 2011: 204); other characteristics include increased lexical innovations, specialized vocabulary, and expansion of sub-genres in scientific writing more generally (Gotti 2011: 209, 218, Culpepper & Kytö 2010: 38-9).

The RS itself was the first "public institution" committed to scientific research (Atkinson 1999: 16, Royal Society 2020) and "saw revolutionary advancements in the conduct and communication of science" (Royal Society 2020). Contradictions were inherent in the beginnings of the Royal Society (Tinniswood 2019, Wottoon 2015). Despite commitments to education and freedom, and the priority of evidence-based theories rather than received knowledge, the RS was a gentlemen's club, an elitist, male-centered group, reflecting cultural norms of the time

(Tinniswood 2019, Wottoon 2015). Although women were not overtly prohibited from joining the Royal Society and various articles were published by women, the first female member was not invited until 1945 (Royal Society 2020). However, as people ceased to "believe in received knowledge" and the institutionalization of science unfolded, "scientific and medical writing became more diversified with the new medium of the printing press" (Pahta and Taavitsainen 2011: xvii).

This work draws on specific background studies in corpus linguistics in addition to historical studies and work on the history of science and medicine. McEnery and Baker note that both the historical context and "an appreciation of language is vital for an understanding of the written records left to us from the past" (2017: ix). McEnery and Baker's work is exemplary in combining corpus-based methods and historical context in their exposition of discourses of seventeenth century prostitution (2017). They utilize keyword and collocations (24-5) to expose changing discourses over time, and collocations evidence "shifts in word meaning and representation" (26).

Additional historical corpora relevant to this work include the *Corpus of Early English Medical Writing 1375–1800* (Taavitsainen et al. 2000), the *Middle English Medical Texts Corpus, the Corpus of English Correspondence*, the *Corpus of Historical American English*, and others. Each of these corpora differ with respect to sampling, time periods, and compilation, but are notable representations of key work in this area of research. Taavitsainen and Phata created the *Corpus of Early Modern Medical English Texts,* a representative sample of medical writing from 1500-1700 (2011) to address the lack of linguistic studies focused on early scientific writing in English (2011: 9-11). Their corpus includes excerpts from the *Philosophical Transactions of the Royal Society,* noting that this journal provides important evidence of the "emerging empirical approach to science" as well as many other genres of texts within this umbrella (2011: 41). This work is utilized to track trends over time and to examine shifts in linguistic patterns in this important and distinctive genre (2011: 16).

Work on the *Corpus of Early English Medical Writing* has emphasized differences between different types of medical texts (Taavitsainen 2011: 110-115), changes in technical and genre-specific vocabulary (Götze 1995, Gotti 2011: 209, 218), and involved studies on specific linguistic devices, like Hiltunen and Tyrkkö's work on verbs of knowing (2011: 172). Verbs of knowing were found to vary across different contexts in Early Modern English (EME) medical writing, with the most variation present at the level of individual texts (172). Taavitsainen analyzes the differences across genres of medical texts, including those addressed to practitioners and to lay people; she notes that "context" is key in analyzing specialized corpora, especially to gain a "multifaceted" analysis of linguistic patterns (95).

Another related corpus study of EME includes Culpepper & Kytö's analysis of speech-related genres in the Corpus of English Dialogues (1560-1760); their findings span pragmatic and grammatical features associated with speech in EME, pointing out the "diversity of influences" on different genres in this time period (2010: 398). They also studied lexical bundles, finding that they were highly utilized as "utterance launchers" (2010: 399). A key aspect of their work involves the consideration of discourse and context on speech communities and situations in EME (2010: 400).

Work on scientific communication has included different genres and historical data sources, although much remains to be done. Previous work in English on science and medicine has found important genre characteristics, including lack of passivization, use of first person pronouns, and increasing technical vocabulary over time (Götze 1995; Gotti 2011: 209, 218; Culpepper & Kyto: 2010: 38-9). In terms of genres and contexts, the scientific journal article as we know it today was not a text type at the creation of the *Phil Trans*. Instead, the epistolary form was the primary context for sharing and spreading scientific knowledge, gradually morphing into the various contexts included in the RSC (Atkinson 1999; Wootton 2015; Taavitsainen and Pahta 2011). In medical writing in EME, Pahta and Taavitsainen argue that "changes in the underlying scientific ideology as well as in the discourse community can be verified both on the micro-level

of individual linguistic features and on the macro-level of argumentative structures and textual organization" (2011: 3).

Additional research has focused on specific linguistic features in historical scientific writing: uncertainty markers and hedges in the British Medical Journal over time (Bongelli et al. 2014; Yang 2013; Oliver del Olmo 2014), modals in the Coruña Corpus of English Philosophy Texts and the Coruña Corpus of English Scientific Writing (Crespo & Moskowich 2016, Monaco 2016, Moskowich et al. 2021), and epistemic claims, modals, and n-grams (Plappert 2017). Oliver del Olmo compares scientific communicative situations between English and Spanish Medical texts and discusses hedging devices, including modal verbs, probability adjectives and adverbs, epistemic verbs, adverbs of quantity, conditional markers, passive voice, and depersonalization (i.e. nominalization and non -personal forms), the use of which has been linked to discourses of communication among medical experts (2014: 275). Similar studies have explored these devices in more modern sources of English scientific writing, including Plappert's account of genetic scientific writing through keyword analysis and analysis of lexical bundles (2017).

Previous work on the *Phil Trans* utilizes a variety of methods, corpus-based and rhetorical analysis (Atkinson 1999, Biber & Gray 2010). These studies have focused predominantly on sub-sections of the *Phil Trans* rather than the publication in its entirety. Findings include the decline of narrative forms and features of epistolaries to more standardized structures and genre characteristics of scientific articles today (Atkinson 1999: xxvi, Gross et al. 2002, Biber & Gray 2010, Biber & Conrad 2019: 241). The compilers of the RSC (Fisher et al. 2020) note that the "differences between early and contemporary scientific articles are profound", and the RSC verifies "expectations with regard to the development of English scientific writing" (2020: 799). Biber and Conrad discuss the development and evolution of the scientific article by

analyzing a subset of *Philosophical Transactions of the Royal Society*[11] (2019). Biber and Conrad note from the beginning of the *Phil Trans* well into the eighteenth century articles were characterized by the "first-hand personal perspective of the author" and expected genre conventions of letters (2019: 237-40). Beginning in the early 1800s, some publications shifted from first-person to "an impersonal presentation of procedures and findings" (Biber & Conrad 2019: 240), and by the end of the 1900s, most publications of the *Phil Trans* concentrated on "theoretical concerns" rather than narratives of experiments. These theoretically focused publications align with current expected conventions of scientific research articles and writing practices today (Biber & Conrad 2019: 240).

Atkinson utilizes methods from rhetorical analysis and multidimensional analysis from corpus linguistics (1999) to analyze only specific and limited sub-sections of the RSC (70 articles) through a sociohistorical perspective, foregrounding the many aspects surrounding the history and development of the Royal Society and the publication of the *Philosophical Transactions*. Even though he discusses linguistic features uncovered through multidimensional analysis (MDA), including use of place adverbials, explicit reference, and use of wh-relative clauses (1999: 110- 140), no other methods from corpus linguistics are utilized in his study of the *RSC*. Atkinson's findings also align with previous literature on the development of scientific writing. He found a decline in references to authors in his examination of sub-sections of the *Phil Trans*, the "replacement of author-centered writing" with "object-centered writing," and the frequent use of stance markers in the 17th and 18th centuries (xxvi).

3. Methods

The Royal Society Corpus (RSC) was made available through the University of Georgia Corpus Server from the Department of Linguistics LingLab. The corpus itself was compiled of

[11] This subset of the *Phil Trans* was included as part of the original composition of the ARCHER corpus, compiled in the 1990s as a representative sample of historical British and American English across genres (Biber and Finegan 1997: 257).

data from the Royal Society and from JSTOR, with analogous metadata including text year, decade, and century. Additional attributes with linguistic information were added, including part-of-speech tags and lemmas (Fischer et al. 2020, Kermes et al. 2015). Text types and author metadata are also included in annotations (Fischer et al. 2020). The articles composing the corpus provide a vast collection of emerging and prominent scientific disciplines over the three centuries, in which "scientific discourse formed as a discipline and underwent considerable changes" (Fischer et al. 2020). The RSC was also developed according to FAIR principles (Wilkinson et al. 2016; Fischer et al. 2020) and is freely available for download and use worldwide.

For analysis of the RSC, R packages are utilized in conjunction with CQP including polmineR (Blaette et al. 2019), tidytext (Silge & Robinson 2016), tidyverse (Wickham et al. 2019), and ggplot (Wickham, Navarro, & Pederson 2021) for visualizations of corpus data. This study utilizes historical findings in conjunction with the analysis of concordance lines (KWIC) and collocations to discern the contexts of specific keywords. Analysis of concordance lines provides a way for researchers to "examine every occurrence of a word or phrase in context" (Baker et al. 2019: 31) and is an advantageous method for finding linguistic "patterns that might be less obvious during more linear, left to right readings of the data" (Sinclair 2003, as quoted in Baker et al. 2019: 31).

Collocations are calculated utilizing the cooccurrences function in polmineR (Blaette et al. 2019), with default settings of five tokens on the right and left of the node. Collocates refer to words that appear as neighbors of other words; they are widely recognized in corpus linguistics for understanding "how meaning is acquired through repeated uses of language" and for exposing discourses in different genres and texts (Baker 2010: 13). Log-likelihood (G2, keyness), raw frequencies, and ratios of use per 100,000 tokens are reported (Dunning 1993, Gabrielatos & Baker 2008). Brezina notes that log-likelihood is especially pertinent for highlighting typical contexts of use (2018).

The selected tokens for analysis were each chosen due to their relation to changing communication around medicine and science and corresponding updates in related technology. They all occur at varying frequencies across the entire duration of the corpus and in each scale of analysis. Different scales are employed for analyzing these data in order to garner a thorough understanding of contexts, use, and specific patterns employed in communication in the *Phil Trans* over time. These scales include the level of text century (1600, 1700, 1800, and 1900), text period divisions (1650, 1700, 1750, 1800, 1850, and 1900), and differing levels of analysis in terms of the frequency distributions, collocations, and KWIC lines at the individual contextual level. These are constructive advantages offered by corpus methods, so that an aggregate view of the corpus at different levels of scale is considered in the overall interpretation, highlighting the diverse and dynamic nature of communication.

4. Analysis and Discussion

The following analysis highlights the fraught nature of trends in significant medical topics and also illustrates different methods for handling smaller amounts of data in historical corpora. *Health, disease(s),* and *inoculate*|*inoculation* are examined via keyword-in-context analysis and collocations. Collocates and concordance (KWIC) lines are obtained utilizing polmineR (Blaette et al. 2019) functions, and different text periods are included with each method[12]. KWIC lines are examined according to text period divisions (50 years), while collocations and dispersion are considered over the entire corpus and by text century. In some situations, the node word is utilized at similar rates per text period, while others display considerable differences in dispersion over time. The rates of usage across these divisions provide interesting evidence of changes in communication; they also are linked to specific meanings and discourse shifts in the semantic prosody of each token over time.

[12] Spelling variants and different capitalizations are accounted for utilizing polmineR functions.

Porter argues that "the triumphs and trials of modern medicine can be understood only in a historical framework" (2006**:** 8), and "medicine has been constantly remaking itself…of course, medicine has always been about the same thing: healing the sick. But what that has entailed- imaginatively, organizationally, scientifically, humanely- has forever been in a state of transformation" (2006: 9). Health and medical developments over the time span of the publication of the *Phil Trans* were far-reaching (Berridge 2016: 4; Porter 2006a; Kusukawa 2004; Wear 2000). Galenic and Hippocratic medicine was widely practiced at the beginning of the *Phil Trans* (Ackerknecht 1982: 98-100; Kusukawa 2004; Conrad et al. 1995; Berridge 2016: 32); this included prioritizing balancing the "four humors" of blood, phlegm, yellow bile, and black bile, "which were constitutive of health and disease" (Berridge 2016: 31; Reiser 1978: 8). Interventions and practice focused on the individual patient (Reiser 1978: 8-9) and upheld that "all aspects of the person were interlinked" (Porter 2006c: 80)**.** The Hippocratic focus on the patient is still considered to influence modern medical practice today (Kusukawa 2004; Berridge 2016).

Overall definitions of health and public health are "subject to endless debate and redefinition" (Berridge 2016: 8). This is highlighted through the ways that medical and societal developments played out across the time span of the *Phil Trans,* including the influence of science and technology. As shown in the corpus data, *health* KWIC lines in the *Phil Trans* reveal the multiplicity of impacts surrounding this term and ongoing changes in connection with modern medical practices. The integration of morality with health is revealed in earlier periods, with discussion related to balancing the humors. Although early medicine "tended to ignore the concept of contagion" (Shyrock 1947: 75), later KWIC lines evidence changing concerns over time including a focus on community-oriented medical approaches, the application of statistics

widely to medical and other sciences[13], and interest in public health (Berridge et al. 2011: 29-40, 47; Berridge 2016: 40-41).

Other scientific influences abound, including diverse findings, the rise of laboratories and new experimental practices to test and confirm hypotheses, and the emphasis on hospitals for the care of the unwell. Discoveries throughout this time period led to new methods for diagnosis. As Porter points out "a sound anatomical and physiological basis is to us essential to scientific medicine" yet this could only flourish from "systematic dissection"[14] (2006c: 137). By the 1700s, innovations "in gross anatomy…and physiology also had created the dream of a scientific understanding of the body's structures and functions, drawing on and matching those of the new and highly prestigious mechanics and mathematics" (Porter 2006d: 142).

4.1 Analysis of patterns surrounding Health

KWIC contexts of the token *health* in the 1650s reveal connections to Galenic medical practices and cultural ideas, including mentions of humors and morality. *Health* is often connected with morality in the Early Modern period, not simply because of the influence of the Church, but also because of the idea that individual choices or sins produce illness (Kusukawa 2004: 19, Berridge 2016: 36). Connecting health and morality provided a framework for practitioners and the public alike to understand and frame diseases and impacts on health (Porter 2006c: 93, Kusukawa 2004: 18-20): "moralizing with sickness…has a *prima facia* appeal…it rationalizes menacing maladies and renders adversity less mysterious" (Porter 2006c: 93).

Other cultural norms are present in this time, aligned with the male-centric founding members of the RS, exhibited in KWIC lines: *life and health of man*, *relief of man's health,* and

[13] Berridge points out that "the growing interest in numerical calculations prompted the gathering of statistical information" implemented by European governments in places including the Italian city states and in France, most prevalent in the mid 1700s (Berridge 2016: 40, Wootton 2015: 262). In the UK, several RS fellows applied mathematics and calculations to study "the conditions under which prosperity flourished and was impeded" and to study statistics regarding diseases and mortality (Berridge 2016: 40-41, Wootton 2015: 263).

[14] Porter notes further the influence of the Church in this area, pointing out that "ecclesiastical opposition to dissection slowly melted away in medieval times" (2006: 137).

the preservation and restoration of the ingenious Author. There is also a focus on health in terms of the body and "spirit" as in *hunting preserves bodily health and alacrity in our Spirits.* This corresponds with previous work on Early Modern English (EME) communication and views surrounding anatomical knowledge and practice; "medical theory and accounts of what patients experienced were relatively prominent in Early Modern Medicine and often merged with anatomical signposting" (Wear 2000: 125). Related terminology in medical theories at this time included terms like "reason, humours, spirits, natural, animal" (Burton 1621: 127, as quoted in Wear 2000: 125). This concern with the *spirit* also connects with Galenic medicine, as Galenic practitioners argued that "human beings carried out three functions of respiration, nutrition, and animation by means of three medical spirits- vital, natural, and animal" (Kusukawa 2004: 6). Early Modern concerns with the spirit and the body were generally based on "scripture, quotidian experience, or sophisticated conceptions of the interchange between matter and spirit. However, the precise interaction between passions, humors, elements, and spirits is complex, often vaguely characterized, and frequently revised in the period" (Pender 2010: 210).

Early Modern attitudes towards the physical body were complex and entrenched in ideas discussed above, including the connection between the spirit, soul, and body, and a distrust of the passions and flesh of the body. At the same time, cultural practices also prioritized the care of the physical body (Porter 2006c: 82-3, Porter 2006d: 150). In the 1700s, KWIC lines highlight these ideas, with many mentions of the *body* in the following lines: *of the human Body in Health in Winter, more conducive to the Health of the Human Body*. *Health* remains connected to humoral theories, as in the 1650s, represented in lines like *the Air, Winds, Health, Fruitfulness*. Humoral theory and balance are connected in terms of different elements and constitutions; for EM medicine, the goal was to achieve balance of the humors to bring about health and well-being (Kusukawa 2004: 17-18). Medical practitioners at this time "focused on using outward signs" to interpret potential internal problems (Kipple 2006: 44). Connections to morality, including a distrust of different mental states and the mind persist in this time period. KWIC lines revealing

this are exemplified in the following: *the Mind proving destructive to health,* and *temperate people* are described as being *in perfect Health. Health* is also connected to larger communities and groups of people in the lines: *serviceable to the Health of Mankind, Health and Trade of Mankind,* and *living Bodies in Health* and *Sickness.*

Discussions around diets and health are present in the lines: *palate rather than the health of the body, liquor every day in Health, Food of mankind during Health.* An awareness of the connection of health and sanitation is described in KWIC line: *employed for the Cleanliness and Health of the Skin.* Cultural norms, including superstitions are also represented: *an ill omen of the Health or other misfortunes.* Finally, notions of health and science are broadly connected through the idea of using electricity as a potential "cure" for illness: *if the Electricity restores Health to a sick person.*

The 1750s includes KWIC lines pointing to cultural norms with an increased focus on *blood* in relation to health. This is unsurprising considering William Harvey's discoveries on the circulation of blood[15] and cultural norms of this time period; *blood* was also considered to be the most important of the humors in Galenic medicine (Porter 2006d: 139). Lines that discuss this include *blood of a person in health is completely coagulated, blood of a person's nearly in health.* There is also more consideration of women's health, with a tendency to focus on appearances and mental states: *improvement in her looks and health.* Divisions in diagnosis and practice across genders persisted well past Early Modern times, and some key examples of this include exhibiting symptoms frequently diagnosed as hysteria in women, with other diagnoses in men (De Rinzi 2004: 198-202). Women were also considered to be anatomically inferior to men, with many diverse explanations relating to this[16] (De Rinzi 2004: 197, 201).

[15] Harvey made several key discoveries, including the core operations of the heart and circulation of blood, which some argue is the "greatest medical discovery of all time" because it "introduced the principle of experimentation in medicine" (Friedland 2009); Harvey published his discoveries in Latin first in 1628 and in English in 1653 (Friedland 2009).

[16] One early example of this is the publication entitled *The Suffocation of the Mother* (1603) in which the author argues that "signs of witchcraft were generally produced by a somatic disease..called 'the Mother',

In the 1750s, KWIC lines connecting to cultural influences include *to the gods for your health and preservation, drinking to the health of all the company,* and *God grants me life and health.* Cultural norms from the medieval period through this time upheld the view that ill-health "could come from God, the stars, or nature herself" (Wear 1995: 219). It is notable that less reference occurs to nature or to the body generally in the 1750s. The line *influence of cold upon the health of the inhabitants of London* connects groups of people, places, and collective ideas about health. This is connected to humoral theories and the idea that individuals belonged in certain climates based on the individual's constitution (Wear 1995: 230, Jenner 2004: 287-8). These ideas created positive results, as politicians and others implemented measures involving hygienic practices, "derived from the belief that climate caused disease" (Wear 1995: 230) and to minimize the possibility of disease (Berridge et al. 2011: 27). By the 18th century, "public health was becoming part of the modernizing state concerned with urbanization and economic development" (Berridge 2016: 41), and early efforts such as these, exemplified in concerns with environment and air quality are often considered to be the beginnings of "modern public health" (Berridge et al. 2011: 26-7).

Beginning in the 1800s, a notable increase in discussions of different functions and aspects of the body are present in the discourse: *sensorial functions being consistent with health declining, nerves in the face of health and disease, urine in state of health,* and *respiratory movements in health and disease.* These lines evidence the increasing effects of the introduction of laboratories, experiments, and overlap between science and medicine. Continuing discussions relevant to public health are also depicted in descriptions of new institutions: *The Board of Health with that promptitude, preservation of human life and health in our great cities, influence of cold in the Health of the inhabitants of London.* Berridge et al. note that "a new body of law and

---

an old term for hysteria (Porter 2006c: 77-8). Galenic medicine also upheld the premise "of an intrinsically weak and inferior female body" but was also a "flexible framework" that allowed other learned practitioners to draw "the opposite conclusions" (De Rinzi 2004: 201).

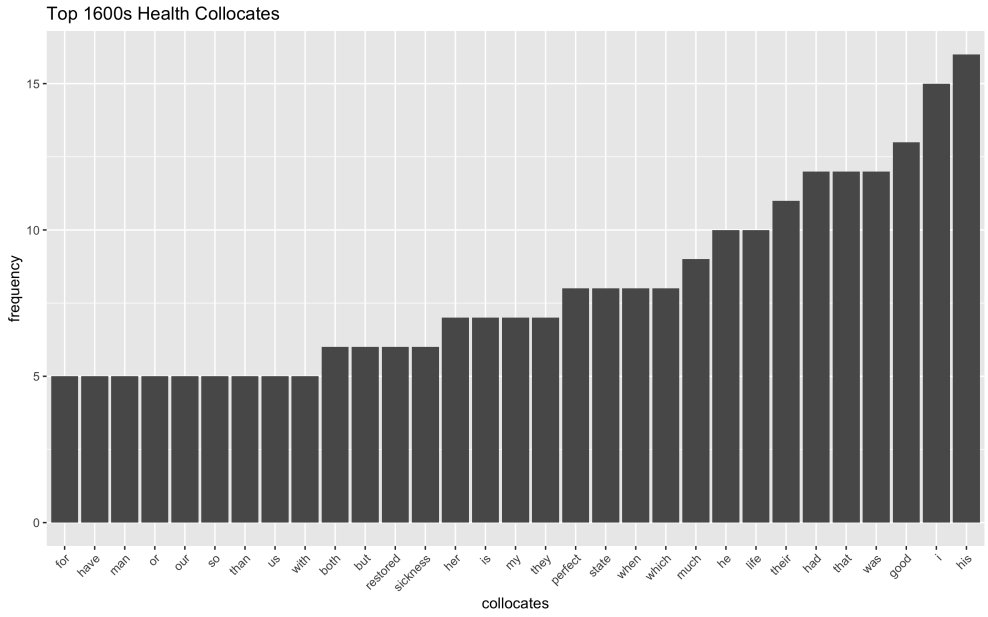
administrative practices" began in the nineteenth century, implementing vast efforts in public health (2011: 27-8). These efforts culminated in scientific discoveries such as Pasteur's Germ Theory (Berridge et al. 2011: 27-8) which "opened up" vast options for understandings of disease, including "surveillance and screening…investigating the safety of food and water" among others (Berridge et al. 2011: 28).

Earlier discoveries culminate in the 1850s through key developments which include direct references to public health: *science of Public Health, ever present dangers to the public health, always with the central Public Health authority of this country, that the science of public health was almost inaugurated in England, Statistical Report of the Health of the Navy, general preservation of public health, collectors of statistics of health's mortality.* Sanitation efforts are discussed in association with health more generally: *subscribed by medical officers of health, sanitary engineers; sanitary functions alike of health office and of local authority; at work to promote health such as improved sanitation.* Sanitation, health, and questions of safety behaviors were considered following discoveries in the mid 1800s, so that "the concern of public health was moving away from collective management of the environment to interventions targeted at individuals in the home" (Berridge et al. 2011: 28). The focus on individuals included health education, exemplified in a concern with exercise, discussed in the following KWIC lines: *under ordinary circumstances of health, slow walking exercise to health for three days.* Specific organs and aspects of the body are linked to health in exemplary lines like *investigation of the organs in Health and disease, the knee jerk varies in Health, his bodily health was feeble,* and *heart's power in Health*.

The last time period in the 1900s underscores the introduction of technical terms and practices of medicine. *Health* is considered in terms of generalized groups: *Infectious Diseases's public Health, with results to the public health which could scarcely have been, that Board relating to Public Health on the Lords, especially in relationship to public health work.* Health is also discussed in terms of the individual and medical practices: *sunlight to influence the nervous*

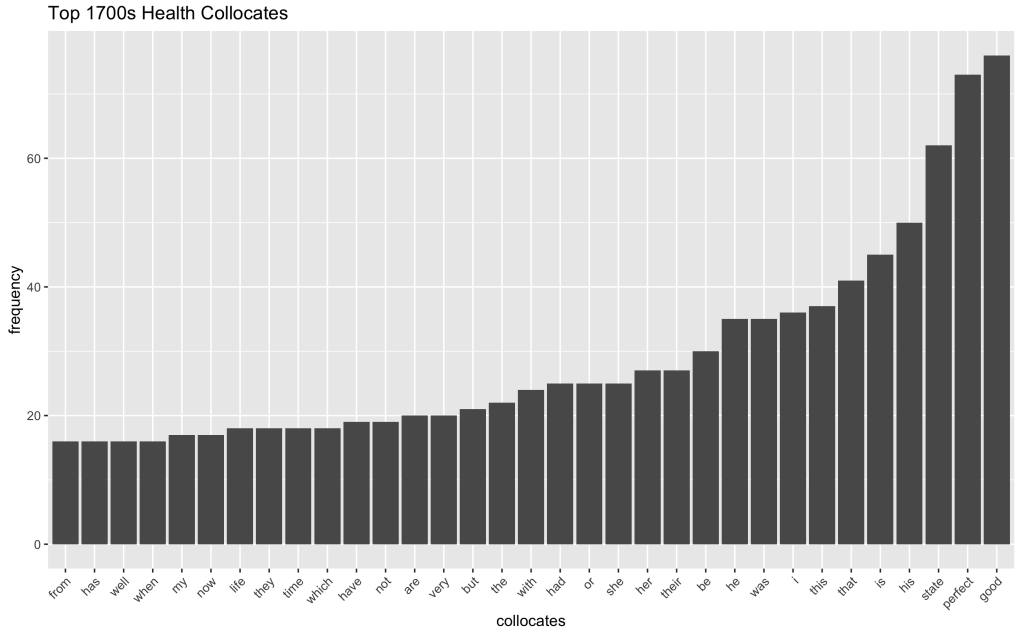
health and metabolism of man, tubercular individual when in normal health and at physiological rest, cystitis and disturbance of health, sphyganographic records taken in health and disease, the tables give the relations health and ability, body weight exceeds the limits of health. This time period involves a greater concern with individuals and health education (Berridge et al. 2011: 27). Many diverse medical and scientific specialties flourished and resulted in the implementation of wide-ranging advancements (Ackernecht 1968: 208-9). Some examples include the chemical analysis of food and diet, "bacteriology" shaping general and pediatric practices, and histology innovations improving knowledge "of the hematology of childhood" (Ackernecht 1968: 201).

Figure 4.1



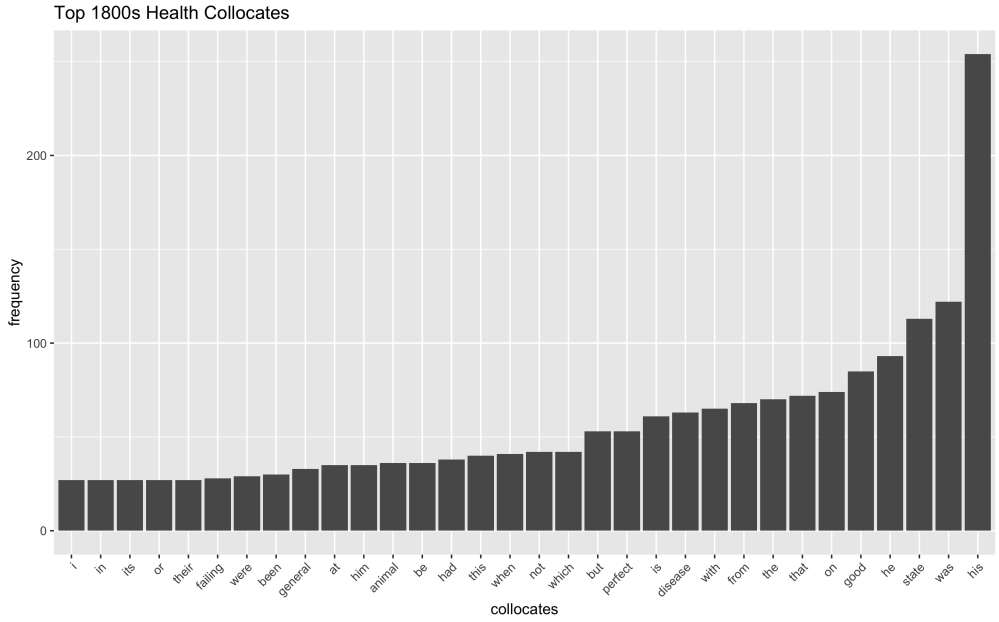
Collocations of health over the entirety of the corpus illustrate the predominant ways of discussing health over time. The top collocates of *health* in the 1600s include *his, I, good, had, their, life, he, state, perfect,* and *they.* A focus on individuals is present, with top collocates including personal pronouns like *his* and *I. Good* is also the most prevalent descriptor of health in the collocational distribution, with others present including *life, perfect, state,* and *restored.* A focus on ill health is represented in small quantities overall through tokens like *diseases, curing, preserving,* and *recovery.*

Figure 4.2



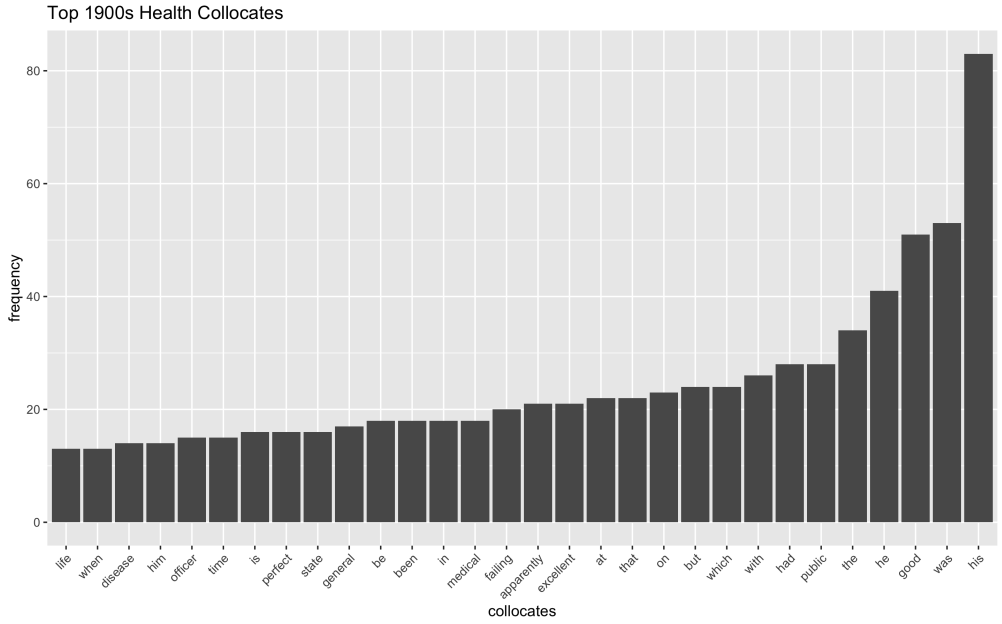
The distributional profile for the collocations of *health* changes in this time period, with general descriptive tokens rising in frequency like *good, perfect,* and *state.* Personal pronouns remain highly frequent, although *his* replaces *I* as the most frequently occurring pronoun. The 1700s top collocates overall are *good, perfect, state, his, that, this, I, he, their,* and *her.* Additional key collocates of note from this century are *time, she, life, my, preservation,* and *body.* Notably, there are no highly frequent collocates that connect health with medical-related terms beyond general descriptors and references to individuals.

Figure 4.3



Collocates of *health* reflect core changes culminating in this time period, including terms like *disease* rising greatly in frequency. 1800s top collocates include *his, state, he, good, disease, perfect, this, animal, him,* and *general. Failing, I, excellent, report, navy,* and *public* are other important top collocates, displaying growing interest in collective concerns and public health, as well as continued use of first-person pronouns and more general descriptors of health. These collocates along with others like *urine, statistical, medical,* and *illustration* also display the growing connection and emphasis on scientific practices and health in medicine.

Figure 4.4



The last century under consideration includes the following top collocates *his, good, he, public, apparently, excellent, failing, medical, general,* and *perfect. State, officer, time, disease,* are other top collocates reflecting key changes at this time. More general descriptors remain in the collocational distribution like *good, excellent, failing,* and *perfect.* This frequency distribution includes persistent characteristics and trends, but the frequencies of many personal pronouns decrease, with the exception of *his* and *he.* Table 1 displays the changes over time in dispersion of the token *health,* with frequencies declining over each century.

Table 4.1: *Health* Dispersion across Text Century

Text Century	**Raw Frequency**	**Frequency per 100,000**
1600	129	5
1700	474	4.86
1800	1100	2.4
1900	421	2.1

The dispersion of *health* across all four centuries of the *Phil Trans* illustrates continuing changes in use. Health is utilized most often in the 1600s but declines in frequency overall until the last time period. This is most likely attributable to changes in genre norms in addition to the shifting focus of the journal itself, including increased knowledge related to health and medicine as diverse scientific disciplines. Regardless, it is of interest that the dispersion of *health* declines so dramatically across each century and reflective of broad changes related to medicine becoming a modern medical practice.

4.2 Analysis of patterns surrounding Disease(s)

While linguistic patterns surrounding *Health* provide one viewpoint to consider discourses in medicine over time, tokens surrounding *disease(s)* provide another valuable vantage point. *Disease(s)* collocates often with *health* across the writings in the *Phil Trans.* Findings for *disease(s)* reveal similar developments discussed above alongside the movement toward scientific classification of diseases, understanding anatomical and physiological frameworks for disease, a focus on individual patients and symptoms, and the development of public health and implementation of modern medical practices. Integral and systematic improvements in knowledge of disease involved the change from a diagnostic heuristic based on presenting symptoms to "understanding the disease" itself (Jenner 2004: 296). Important contributions to understandings of disease occurred prior to the nineteenth century, but "systematic epidemiological and pathological research programmes did not develop" until that time period (Porter 2006d: 150). Illnesses were generally attributed to personal decisions, including explanations related to morality (Porter 2006a, Kusukawa 2004, Berridge 2016), and other theories related to contagion development across this time. Around the turn of the nineteenth century, increased movement of the sick into hospitals in Western Europe and in England was distinctive and impactful. This development is seen in several key ways, including a separation of the sick and the well; hospitals in the UK became integral for uniting medicine and science (Porter 2006d: 156).

In the 1650s, *disease(s)* is used in many contexts discussing specific individuals: *Queen Mother was troubled with this disease* and people more generally: *people that die of that Disease.* This focus on the individual is in keeping with Galenic and medical practices in Early Modern England. Practitioners also focused on exhibiting signs in the body to understand diseases, and KWIC contexts also discuss different symptoms and names of diseases. Some distinct examples are: *Arteries in the Tooth-ache a Disease analogous to Gout, cure to so stubborn a Disease, dying of the Disease called Hydrophobia, an Apoplexy ensues which disease will also dangerously happen,* and *her falling Sickness a disease never without the suspicion.* The last line overlaps with problems concerning diagnosis across genders; various illnesses received different diagnoses depending on whether the patient was male or female[17] (De Rinzi 2004: 198-202). KWIC lines also illustrate concerns with different aspects of diseases: *gives a History of the Disease.* This time period included important developments involving scientific studies interested in applications of physics to understanding the body as a mechanical object (Porter 2006d: 142).

In the 1700s, *disease(s)* is often connected to Nature: *Nature or from a Disease or provoked by Art, Nature and Stages of the Disease.* This connection of nature and disease is a hallmark of classical medicine, as nature was considered to heal and support bodily processes (Porter 2006d: 142). Discussions are also considered in terms of additional sources of knowledge: *Helvetius*[18] *no ways mentions the Disease* and to EM medical practices: *this particular State of the Disease anciently put the Surgeons, a stricter Inquiry into the Disease by Dissections.* Symptoms and cures are also contextualized, including diet and use of medicines and names of particular symptoms: *very probably that this Contagious Disease, on the whole Disease the Pustules came, the first symptoms of the Disease either by Astringent Medicines, the Method of Curing a Disease.* Particular diseases are also included in RSC discussions, including venereal

[17] Another example of this is the widely held idea that tuberculosis was "primarily a female disease" in the 1800s (Porter 2006c: 92).
[18] Helvetius was a French Enlightenment philosopher.

disease (*the Antiquity of the Venereal Disease long before the Discovery)*, leprosy (*Leprous People but that Disease being ceased*), and small-pox (*There is not perhaps any disease more fatal than the Small-pox*).

In the 1750s *disease(s)* occurs in a variety of contexts, with KWIC lines identifying and considering knowledge relating to diseases generally: *mentioned in the histories of diseases, communicate their observations upon these diseases, considering the nature of the disease as uncertain.* Contexts also refer to different symptoms: *the first attack of the disease, a slight pain; symptoms resulting from the different diseases; violent convulsive disease.* People are categorized in different ways relating to disease: *Age, Sex, and Disease, and among the whites the disease shows itself at the beginning* and references to specific locations and groups of people with diseases: *Observations on the Population and Diseases of Chester, Diseases of the Army, Disease of Minorca.* Although various methods for categorization of diseases occurred from the beginning of the *Phil Trans,* these lines reflect a growing interest in the quantification of diseases and different areas of the sciences more broadly (Berridge 2016: 40, Wootton 2015: 262-3). Specific areas of the body are discussed: *An Account of an extraordinary Disease of the Skin, most cutaneous diseases, diseases of the neck, colicky and icteric diseases often arise from gall-stones.* Other contexts refer to patients and to potential treatments and remedies.

In the 1800s *disease(s)* occurs in contexts reflecting growing changes in science and medicine, with an increase in mention of medical instruments like the *catheter.* The 1800s included many integral technological developments like the invention of the kymograph and advances in the capabilities of the microscope (Porter 2006d: 159). References to specific case studies proliferate in this text period, pointing to the growing influence of scientific methodologies and studies on medicine. A concern with the environment and particular locations of disease increases: *in countries where this disease is liable to occur; extensively epidemical disease of our climate; examined at Plymouth, free from disease.* This is connected to growing improvements to epidemiology; scientists were gradually moving away from miasmatic theories

of tropical disease (Porter 2006d: 163-5). Specific anatomical areas of the body are described in reference to *disease(s)*: *the oblique muscles in diseases, nerves thus detected by diseases, diseases of the arteries, kidneys, Animalcules*[19] *has no reference to disease of the skin.* References to *sympathy* in relation to disease are present: *this sympathy from disease, two eyes sympathise in disease.* Symptoms are considered: *acute inflammation in chronic disease,* and mental health-related concerns are present in the line *the hydrophobia is a disease of the sensitive.*

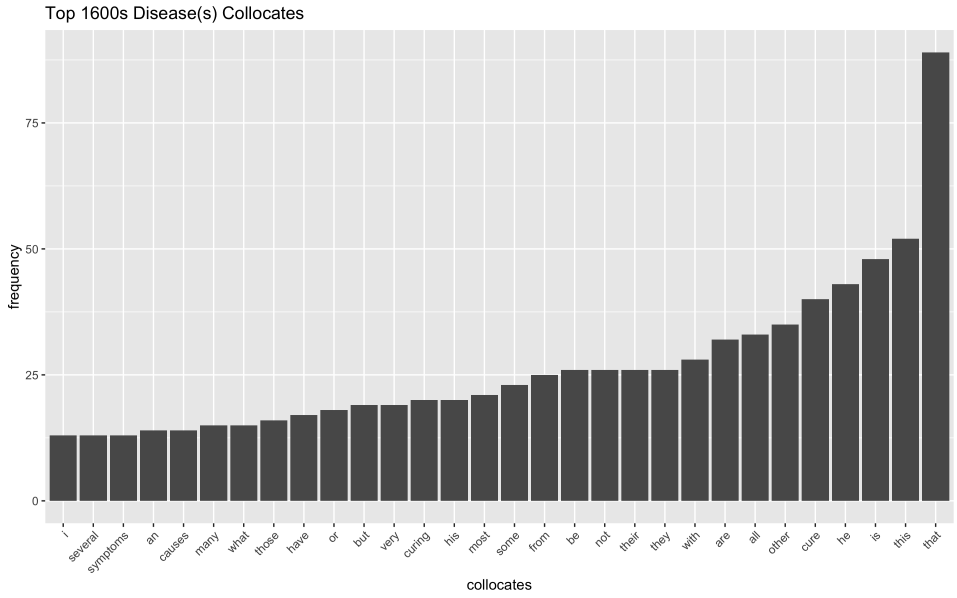
In the 1850s *disease(s)* occurs in increased references to the environment, including concerns with sanitation: *sewage, sewage Produce, and Disease, rapidity with which epidemic disease acts upon different people, connexion between disease and impure water, ravaged by war and disease.* New theories are referenced: *on the Germ-Theory of Disease, parasitic Theory of the disease, when the Germ Theory of Disease was a speculation.* A rise in new technical and medical vocabulary is also present: *Anthrax or Splenic fever, bacillary disease, tuberculosis disease, epidemic fungus disease,* including references to specific body parts and anatomical systems: *heart disease* (*chiefly valvular*), *paraplegia due probably to disease of the dorsal region, diagnosis and localization of cerebral disease, gloso-laryngeal paralysis.* Porter notes that the "enshrinement of physiology as a high-status discipline was a key feature of the 19th century medical science" (2006: 158), and these KWIC patterns evidence growing interests and developments in these arenas. The use of laboratories and hospitals for evidence and experimentation hallmarked these developments (Porter 2006d: 157) and changes displayed in linguistic patterns surrounding *disease(s).*

In the 1900s, *disease(s)* occurs in the context of increased technical vocabulary and references to specific diseases: *prognosis in heart disease, Hodgkin's disease, in this disease as in malaria, chronic trypanosome disease, zymotic diseases, immunization against disease.* References to public health concerns are also present: *the epidemic spreading form of the disease,*

[19] The term *animalcules* was coined by Dutch natural philosopher Antonie van Leeuwnhoek in 1677; it was used to describe very small organisms (Lane 2015).

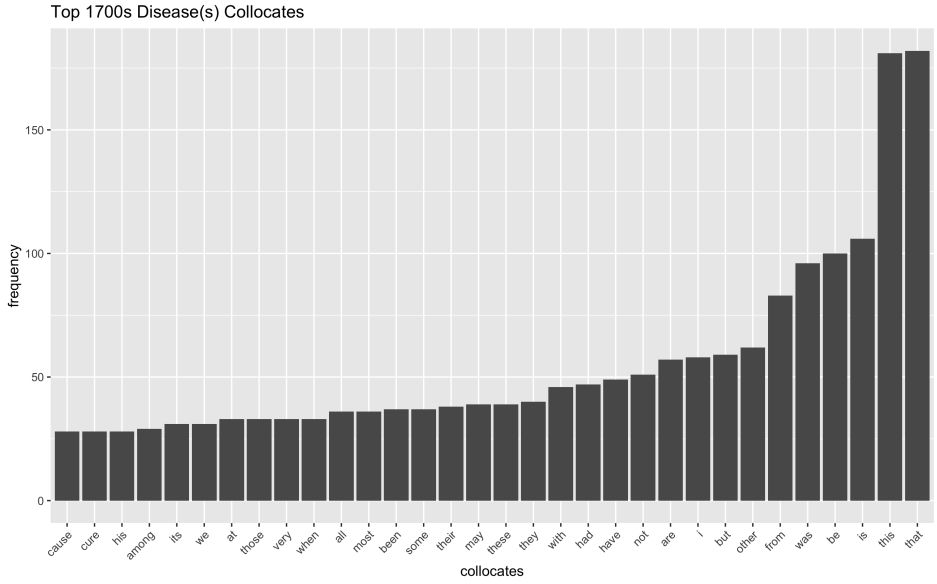
effect of cold on disease, *strangely mysterious and deadly diseases of tropical countries, preventives and remedies of tropical diseases.* Important developments occurred which include the "idea of deficiency disease- the notion that a healthy diet required very specific chemical components" (Porter 2006d: 168).

Figure 4.5



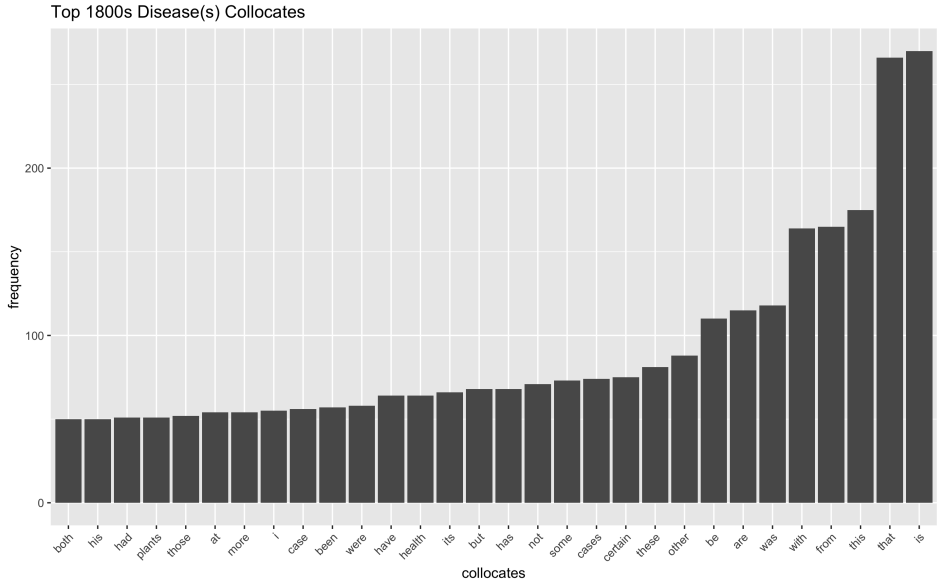
Collocations for *diseases* are separated by text century. The top ten collocations for the 1600s include *this, he, cure, other, all, they, their, most, curing,* and *his.* Additional top collocates in the frequency distribution are *many, very, causes, symptoms, several,* and *I.* These collocates reflect findings from KWIC lines including a focus on individuals and signs of diseases. The prevalent use of personal pronouns is also expected following the early *Philosophical Transactions* epistolary form genre conventions.

Figure 4.6



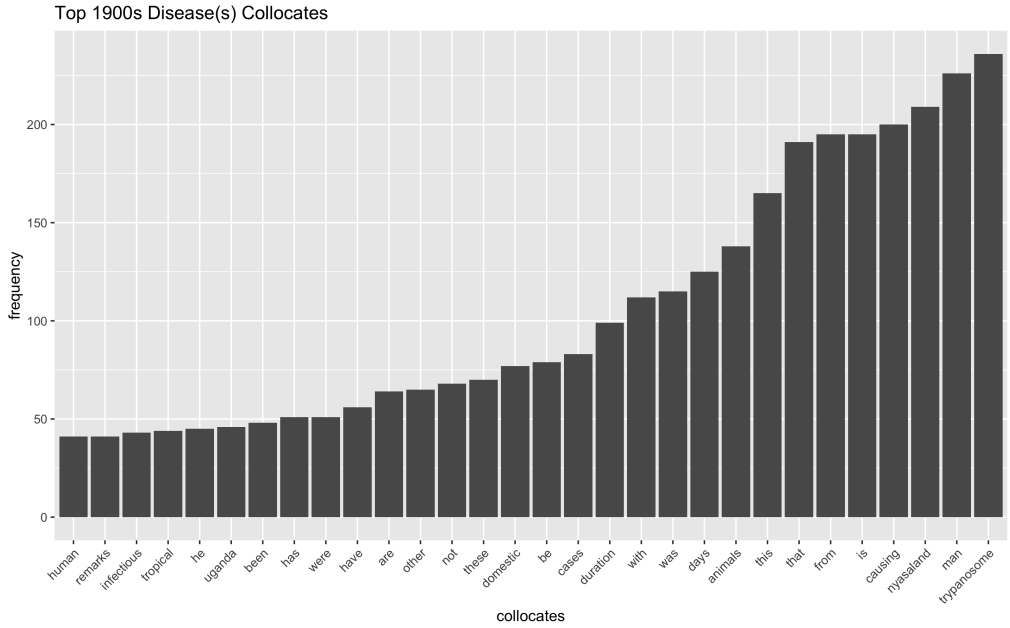
In the 1700s, the collocational distribution changes; top tokens include *that, this, other, I, have, they, may, their, all, when,* and *we. Cure* and *cause* are also top collocates at this time, and first person *I* rises in collocational frequency in comparison with the sixteenth century data. Notable top collocates with *disease(s)* in this time also include *different, venereal, he, putrid, nature, symptoms, causes,* and *fatal.* These collocates emphasize concerns with specific diseases and additional methods of describing and understanding them.

Figure 4.7



There are key changes occurring in the 1800s. The top collocations are *this, other, these, certain, cases, health, case, I, more, plants, his.* Additional top collocates of interest are *heart, symptoms,* and *he.* Other medical-related collocates are *acute, infectious, chronic, nervous,* and *treatment*, signaling increased technical vocabulary and descriptions of disease. These collocations match the KWIC lines from the nineteenth century corresponding with facilitation of diverse methods and practices like experiments and case studies described in the *Phil Trans* in this time period.

Figure 4.8



The nineteenth century collocates reflect crucial developments, including growing definitions of *health* and *disease,* attention to additional symptoms and specific anatomical areas of disease, and a tendency to focus on males. The last century includes the following top collocates: *trypanosome, man, Nyasaland, causing, that, this, animals, days, duration,* and *cases. Domestic, Uganda, tropical, infectious, human, malignant,* and *treatment* are additional top collocates of interest that reflect changes in the 20th century and different areas of concern with *diseases*. The following table shows the dispersion of these terms across each century.

Table 4.2: *Disease(s)* Dispersion across Text Century

Text Century	**Raw Frequency**	**Frequency per 100,000**
1600	499	19.3
1700	1091	11.18
1800	2414	5.24
1900	2168	10.75

The overall dispersion of *disease(s)* differs from the dispersion of *health.* It occurs most frequently in the 1600s, and interestingly, it is least frequent in the nineteenth century. In the last time period, *disease(s)* rises again in use. The different sizes of subcorpora could be related to

frequency changes, as the nineteenth century is by far the largest in size and most prolific. It is also reflective of the fluctuations in concern and focus on different topics like *health* and *disease(s)* across the publication of the *Philosophical Transactions.*

4.3 Analysis of patterns surrounding Inoculate|Inoculation

In order to understand the communication surrounding key technological, medical and scientific developments, the example of the practice of inoculation is chosen, and KWIC lines and collocations are examined over time for the tokens *inoculate|inoculation*. "The introduction and diffusion of inoculation involved complex processes of cultural exchange and negotiation" (Jenner 2004: 331). The practices involved in inoculation were highly controversial and problematic, evolving greatly over time (Jenner 2004: 302-4). In many early instances, the people involved died, and there were no agreed upon norms of methodology; in general, inoculating individuals involved insertion of smallpox pustules into a cut or open wound (Jenner 2004: 301; Kipple 2006: 33). The practice of inoculation "even resulted in epidemics, but after the 1760s safer inoculation methods were found," especially after Jenner introduced the cowpox vaccination (Kipple 2006: 33). In England, *inoculation* was widely publicized through the *Phil Trans,* the first English writings on the topic (Royal Society 2020) and practiced throughout elite society, including by Lady Mary Worthey Montague, and later received the "royal seal of approval when the Prince and Princess of Wales had their two children inoculated in 1722" (Jenner 2004: 302). Despite this, throughout Western Europe there were major divisions regarding inoculation, and the dramatic variation in use of tokens surrounding *inoculation* in the RSC reflect this.

In the 1650s, *inoculation* and *inoculate* occur at very low frequencies (only sixteen times total); contexts reveal concerns with safety and methods: *the fittest tire* [time] *to inoculate is presently after mid-summer, opinion concerning that manner of Inoculation, quickest and safest way of inoculation, it seems that inoculation will hold best and longest.* Other descriptions include *curious in inoculation, season of inoculation,* and noting that there are *very many to*

inoculate. These descriptions provide a baseline of understandings and lack of systematic practice of inoculation with its early introduction and use in Western Europe and worldwide.

Beginning in the 1700s, *inoculate* and *inoculation* occur more often, at a total of 51 times. It is discussed in the context of *Small Pox* and symptoms co-occurring with *inoculation*. KWIC lines reveal the contested nature of inoculation in descriptions like *hazard of inoculation, the present Disputes about Inoculation, this Number of Opposers of Inoculation affirm, practise this Method of Inoculation, Account of the Success of Inoculation,* and *Inoculation is a sufficient Preservative.* These underscore the contested and controversial nature of inoculation in the eighteenth century. Results are also discussed in the *Phil Trans*, including unfavorable outcomes: *second Person that died after Inoculation, fifth that died upon Inoculation was a Woman Servant, see any ill Effects of Inoculation, received the Small Pox by Inoculation, the Practice of Inoculation tends to the Preservation of, that Distemper raised by Inoculation is really the Small Pox.*

Beginning in the 1750s, *inoculation* is described in terms of different locations where it is practiced, including *Holland*, *Boston*, *Denmark*, *Syria* and *Palestine*, the *East*, *Africa*, *America*, and *Georgia*. It occurs a total of 136 times in this text period and is referred to increasingly as the *practice of inoculation, public or private practice of inoculation.* Contexts are more positive than in earlier time periods, with KWIC lines including results like *then undoubtedly inoculation will be recommendable, confirm the utility of inoculation, inoculation is approached, recommended, inoculation succeeded, the success attending inoculation after much opposition.* Even so, negative contexts are also present: *that the practice of inoculation may not be tolerated, the law prohibiting inoculation, prejudices and objections made against inoculation,* and *such as are averse to inoculation have obstinately refused to acknowledge.*

In the 1800s *inoculate* and *inoculation* occur only three times total, with general references to the effects of inoculation, and its positive effects *invaluable present of the vaccine inoculation.* In the 1850s, this token is utilized far more often than any other time period.

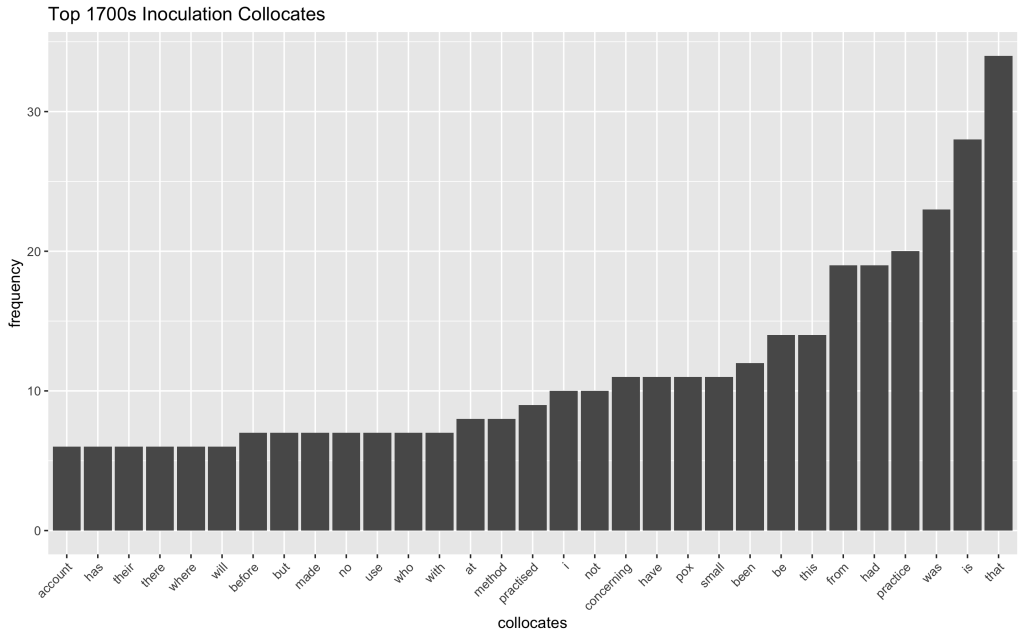
Contexts examine types of inoculation on different animals, as well as humans: *invariably produced by intracranial inoculation, reactionary symptoms after inoculation, Methods of Preparation and Inoculation with Virus, Protective inoculation in the Dog, used in subcutaneous inoculation, with regard to the protective inoculation of man, lymphatic glands immediately after inoculation.* References to vaccination are also present in keyword lines: *of vaccination and of preventive inoculation in combating epidemics.* Negative contexts include *the rapid method of inoculation is dangerous.* Changes in the integration of science and medicine are exemplified in the specific studies discussed in the following KWIC lines: *the study of the anti-cholera inoculation in India, on Preventive Inoculation the tissues, the Plan of anti-plague Inoculation, broth cultures on inoculation into gelatine yielded, on the injecting-syringe when the inoculation with lymph was performed, leucocytosis at the seat of inoculation.* In these text periods and into the 1900s, references to opinions and specific individuals involved in inoculation, as either the subjects or practitioners decline dramatically in use.

The 1900s includes the most references to *inoculation,* with the majority of contexts referring to clinical trials and tests. The trials being discussed are predominantly taking place on rats and animals; however soil, monkeys, cattle, and other organisms are considered. Inoculation results on individual patients are present in KWIC lines, although these decline greatly in frequency: *the patient was reinoculated, Series of Patients before inoculation, case of Patients Undergoing Anti-Tubercle Inoculation.* No mention of negative opinions regarding inoculation or negative outcomes of inoculation occurs in this time period; instead, the authors in the 1900s focus on clinical trials, solutions, methods, and results of inoculation at different points in time: *from this culture inoculation was made into beef-broth agar, after inoculation with oxidized ammonia solutions,* and *after cleansing the inoculation area*.

The tokens *inoculate* and *inoculation* vary greatly in use over the entirety of the corpus. Collocations include a variety of tokens characterizing communication in the RSC. Top collocations of *inoculation* include *season, made, seems, self, who* in the 1600s; these are all at

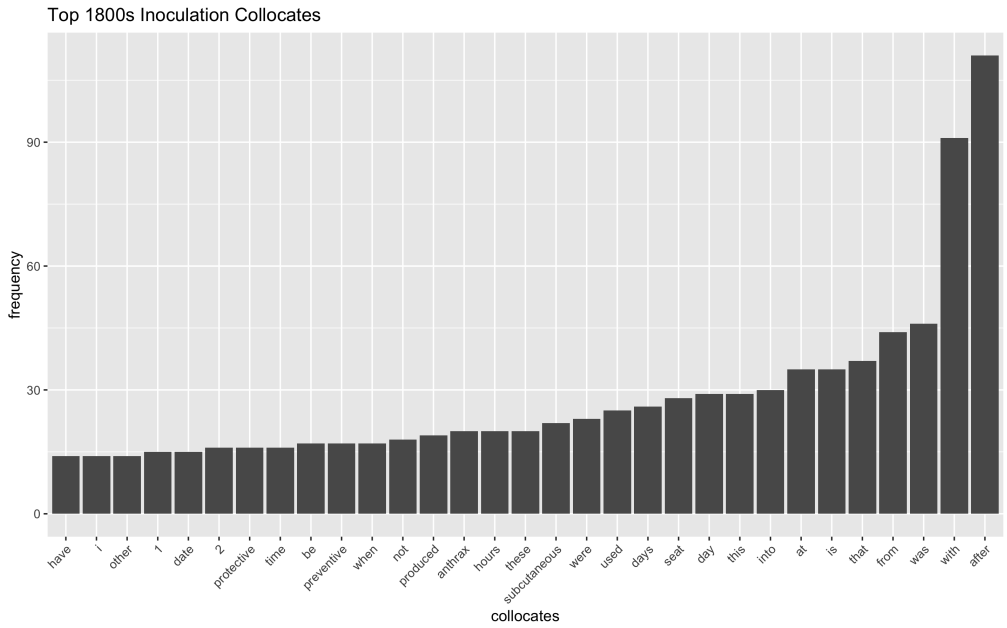
quite low frequencies because of the limited number of occurrences of these terms in the seventeenth century. Even so, collocational results emphasize KWIC findings with corresponding collocates discussed above. Due to these low frequencies, the distributional profiles of these collocations are depicted over time beginning in the eighteenth century.

Figure 4.9



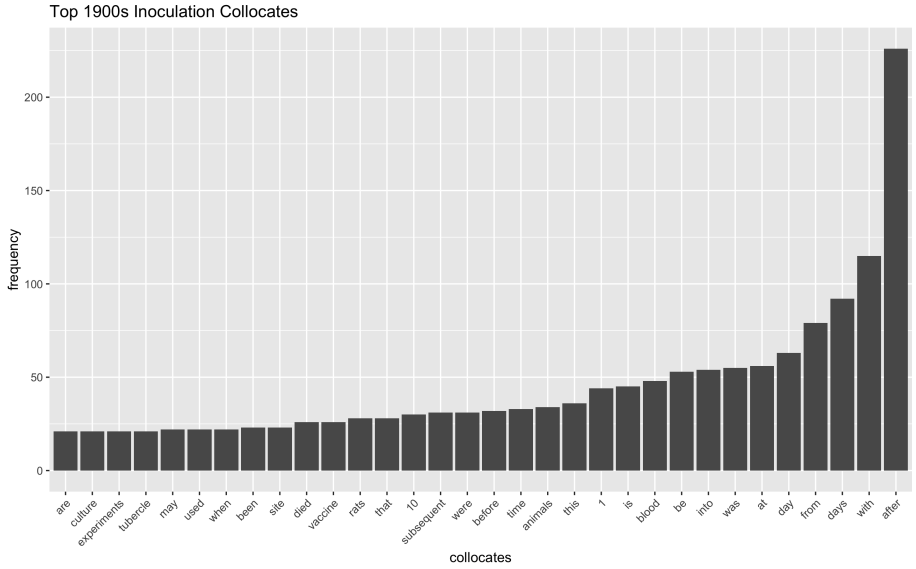
Top collocations in the 1700s include *that, practice, small, pox, concerning, I, practised, method, before,* and *their.* Additional noteworthy collocates are *account, distemper, hazard, received, success, years, application, children, natural,* and *infection.* These underscore the contested nature of inoculation itself and KWIC lines discussing results and symptoms of *inoculation.* Collocates like *method, application, practice,* and *practiced* also reveal the controversial results associated with inoculation.

Figure 4.10



1800s collocates include *after, day*(*s*)*, this, seat, used, subcutaneous, hours, anthrax, produced, preventive. Protective, time, date, I, other, animals, experiments, intracranial,* and *case* are noteworthy collocates that emphasize important differences beginning in the nineteenth century from this time period, like case studies on different animals and attention to detailed methods of inoculation.

Figure 4.11



The top 1900s collocations are *day*(*s*)*, blood, animals, time, subsequent, vaccine, rats, died,* and *site. Experiments, culture, tubercle, hours, cells, result, mice,* and *protective* are other notable top collocates from the twentieth century. These collocates underscore findings from the KWIC lines at this time, with continued advancements in scientific and medical practices. Table 3 displays changes in dispersion in the tokens *inoculate*|*inoculation* across each century of data.

Table 4.3: *Inoculate*|*Inoculation* Dispersion across Text Century

Text Century	**Raw Frequency**	**Frequency per 100,000**
1600	16	.6
1700	187	1.9
1800	515	1.12
1900	853	4.18

Inoculate and *inoculation* increase in frequency over time and across each century. The 1600s include the least frequent examples of these tokens, and the 1700s and 1800s are at close to the same frequencies of use. The dispersion underscores the increase in varied practices of

inoculation until the 1900s, despite the surrounding disputes from the beginning of its implementation.

5. Conclusion

This chapter demonstrates the substantial changes that the "scientific pursuit of medical knowledge" has undergone across the *Phil Trans.* Linguistic patterns surrounding health, disease, and inoculation evidence key scientific, technological, and cultural developments in a unique and important historical source in the RSC. There is a flow of information development in print communications from Early Modern to Modern networks; this includes rapidly changing and developing institutions, and the interplay of theories and evolving medical practices. The KWIC lines and collocations of each of the above tokens coincide with previous findings on medicine and science, with a focus on different aspects of the body, of individuals in the collective and singular sense, and different aspects of health. The 19th century generally includes the most dramatic developments, with an increase in technical vocabulary, broad focus on public health, and recognizable scientific processes upheld in terms of case studies and experiments.

The "origins of modern science" are intimately connected with developments and efforts in cultural and medical understandings of health and disease (Wootton 2015: 21-22), with implications and influences persisting into our current age. Early Modern notions of health were foundational to other ideas about man, nature, and inherited knowledge, and as scientific theories developed, scientists and medical practitioners began applying those findings to discussions and practice. The *Phil Trans* evidences the consequences of science on medicine and cultural ideas, with discourses pointing to growing concerns with public health and a movement away from humoral, Galenic theories. Wootton argues that "a revolution in ideas requires a revolution in language" and further that "the revolution in language is the best evidence that there really was a revolution in science" (2015: 48); this is also present in the *Phil Trans*. As new theories and technical vocabulary developed generally in science, they were applied to the specific sciences of medicine and anatomy.

This work highlights the necessity of different methods, informed by historical research when analyzing important data, and the need for continued work on scientific and health communication. “Public health cannot be understood without a knowledge of its history” (Berridge 2016), and a major facet of its history is its communication. Additional questions relevant to this work include to what degree the *Phil Trans* was representative of predominant cultural norms at different points in time and widely held views of health and disease over time. More work is also warranted considering the influence of external networks and institutions.

CHAPTER 5

US AND UK PRESS DISCOURSES OF E-CIGARETTES

1. Introduction

E-cigarettes have been the most commonly used tobacco-substitute products by youth since 2014 in the United States according to the American Academy of Pediatrics in a recent report (Walley et al. 2019). Walley et al. further discuss information detailing marketing campaigns specifically directed at youth, leading to increasing sales and growth of e-cigarette companies and increasing numbers of young adult American users (2019). Reports from the UK show similar trends (Royal Society Report). The following paper discusses ongoing research about press descriptions of e-cigarettes in United States and United Kingdom press and demonstrates the relevance of corpus approaches for public health concerns. This data-driven study is based on representative corpora, with data collected from January 2010- March 2020. Each article contained the keyword: *e-cigarette(s)* at least once and was obtained via the University of Georgia Library Databases, specifically from ProQuest and from GaleOneFile: News. The US and UK press sources are listed in Table 1 and include national and regional newspapers. Currently the US corpus is approximately 2.6 million words, and the UK corpus is 1.4 million words. The publications were selected according to the specified time frame and to give a differentiated cross-section of both countries by using a variety of news sources from different geographical locations.

Table 5.1: Press Sources

US Sources

Newspapers, National:	*New York Times,* ProQuest
	The Washington Post, ProQuest
	USA Today, ProQuest
	Wall Street Journal, ProQuest
	Radio (news articles from website), *National:* *NPR*

Newspapers, Regional: *Arizona Republic* [West], ProQuest
Atlanta Journal Constitution [Southeast], ProQuest
Baltimore Sun [East], ProQuest
Boston Globe [East], ProQuest
Chicago Tribune [Central], ProQuest
Detroit Free Press [Central], ProQuest
Journal Record (Hamilton, Alabama) [South], ProQuest
Los Angeles Times [West], ProQuest
Newsday (Long Island, New York) [East], ProQuest
Portland Press Herald [Northwest], ProQuest
Salt Lake Tribune [West], ProQuest, GaleOneFile: News
South Florida Sun-Sentinel [South], ProQuest
St. Louis Post Dispatch [Central], ProQuest
Star Tribune (Chatham, Virginia) [East], ProQuest, GaleOneFile: News
Texas Tribune [West], ProQuest, GaleOneFile: News
The New York Post [East], GaleOneFile: News
The Philadelphia Inquirer [East], ProQuest

UK Sources

Newspapers, National: *The Daily Mail,* ProQuest [tabloid]
The Daily Mirror, ProQuest [tabloid]
The Financial Times, ProQuest [broadsheet]
The Guardian, ProQuest [broadsheet]
The London Evening Standard, GaleOneFile: News
The Mail on Sunday, GaleOneFile: News [tabloid]
The Sun, ProQuest [tabloid]
The Sunday Mirror, ProQuest [tabloid]
The Telegraph, ProQuest [broadsheet]
The Times, ProQuest [broadsheet]
Radio (news articles from website), *National: BBC*

Newspapers, Regional: *Banbury Citizen,* [Southern England], GaleOneFile: News
Edinburgh Evening News, [Edinburgh, Scotland], GaleOneFile: News
Manchester Evening News [England], GaleOneFile: News
The Argus [Brighton, England], GaleOneFile: News
The Belfast Telegraph [Belfast, Northern Ireland], GaleOneFile: News
The Birmingham Post [Central England], GaleOneFile: News
The Herald [Glasgow, Scotland], GaleOneFile: News
Wales on Sunday [Wales], GaleOneFile: News

Previous studies have demonstrated the utility of corpus linguistic applications to tobacco advertising and health communication (Kretzschmar et al. 2004, Hunt & Harvey 2015, Atkins & Harvey 2010), press views of mental and physical health issues (Hannaford 2017, 2019, Potts & Semino 2017, Semino et al. 2015, Whitley & Wang 2017, Bowen 2019), and rhetorical strategies used by news outlets (Baker et al. 2013, McEnery & Baker 2017).

This research project focuses on linguistic and rhetorical strategies in actual use, ongoing cultural discourses, and press discussions and differences between places and spaces over time between the US and the UK. News text representations of e-cigarettes and e-cigarette users exhibit temporal variation and differ by geographic location, nationality, and type of news text. These issues are highly fraught in terms of linguistic questions, rhetorical representations, and public health; and there may be significant differences between the US and UK in this regard. In addition to this, e-cigarettes are controversial; some medical research argues for their benefits in the public health sector, while others attack them, implicating e-cigarettes in adverse health outcomes. E-cigarettes have also risen in popularity in recent years in all age groups and are now in widespread use in both the US and the UK (Royal Society Report, Walley et al. 2019, casaa.org).

2. Related Literature

Corpus linguistics is the foundational background of this work; it employs methods to analyze language in use, with the aid of computer technology (Hunston 2012). Although there are differing views of the exact place of corpus linguistics in the multiplying subfields of linguistics and data-science analyses (Hunston 2012, Gries 2009), a major goal of corpus linguistics is to study language based on real life examples of use (McEnery & Wilson 1996). One of the appeals of corpus linguistics is that it applies to many different linguistic topics as well as interdisciplinary research (Brezina 2018).

Several relevant corpus methods are collocation, frequency analysis, and analysis of n-grams over time. These methods are important for identifying linguistic and rhetorical patterns in corpus data that are often connected with cultural issues (Hunston 2010: 162-166). Collocations involve searching for the most frequent tokens surrounding a node word of interest, while n-grams have been utilized in traditional corpus studies as a variation of types of frequency analyses (Scott and Tribble 2006, Scott 2010: 148, Römer 2010). N-grams provide insight into how genres, text types, and discourses use language structures (Scott & Tribble 2006: 132, Römer 2010).

Media and news texts in particular provide an important data source in that they have the power to 'shape widely shared constructions of reality' (Mautner 2008: 32), and this is accomplished through language (Baker 2013: 3). In general, media have an important impact on readers, whether through a technologically updated medium (the Internet and online) or through print newspapers (Jaspal & Nerlich 2017: 481, Mautner 2008: 31). Print media affect readers, partly through their 'intensity of usage, public attention and political influence' (Mautner 2008: 32, Jaspal & Nerlich 2017: 481). Press reports specifically construct and promote "ideologically motivated representations or versions of reality, aimed at persuading people that certain phenomena are good or bad" (Baker 2013: 3). News texts also reflect 'a social mainstream', ideology, or set of ideologies (Baker 2013: 3, Mautner 2008: 32). These media each come with a certain set of expectations, complete with their own set of 'rules', representations, and characteristics. Even more specifically, news texts take this further because they construct and present 'ideologically motivated representations or versions of reality, aimed at persuading people that certain phenomena are good or bad' (Baker 2013: 3). Indeed, even within this genre, there are different ways of representation accomplished through language; for instance, UK press includes tabloids and broadsheets in both national and regional news outlets. The tabloids have a special style with articles that are likely to be shorter and include puns in the titles and main body of the article text (Baker 2013: 32). Corpus studies have successfully incorporated news texts and different types of media into data-driven research, in a variety of interesting and relevant studies.

Baker notes that readers and press audiences "may not be consciously aware of the ways that certain groups are being positioned via language" (Baker 2010: 314); this could easily be applicable to other types of media. Fairclough emphasizes this further: "the effects of media power are cumulative, working through the repetition of particular ways of handling causality and agency, particular ways of positioning the reader" (1989: 54, as quoted in Baker 2010: 314). Another way of describing this is that "far-reaching political developments were bound to have linguistic consequences" (Mair 2006: 7).

Corpus linguistics has a wide variety of applications for media, news, and press, including the consideration of topics incorporating geography and change over time (Paterson & Gregory 2019, Baker et al. 2013, McEnery & Baker 2017). Important corpus research has focused on using news text as data, such as Baker's study of representations of Muslims over time (2013), McEnery and Baker's study of historical news in *Corpus Linguistics and 17th Century Prostitution,* and on discussions of antisemitism in the news (Partington 2012). Other studies relate specifically to medical and health topics. Some work has focused on the media more generally, such as Ringrow's probing the language of cosmetics and its relationship to issues with gender representations and identity in Western media through discourse analysis (2016). These projects and many others demonstrate that corpus linguistics offers useful insights into repeated patterns of language in the representation of social groups and ideas in newspapers (Gupta 2015: 5). Research on media, especially news texts, has more recently been applied to health-related topics, including a focus on mental and physical health (Hannaford 2017, 2019, Hunt and Brookes 2020), and representations of obesity and disabilities (Baker 2006). Hannaford's work on mental and physical health uses corpus-based methods including keyword analysis and semantic tagging to uncover the differences in representations in press (2018, 2019), and Baker uses similar methods with discourse analysis (2006).

Bowen et al. examine depictions of diabetes and schizophrenia in British tabloids; findings involve confirmation of "graphic and violent language" and describing individuals diagnosed with schizophrenia as "as frighteningly 'other' and prone to violence" (2019: 146-8). Bowen et al. conclude that these depictions and othering may contribute to overall negative stigma and stereotyping of individuals who experience psychosis (2019: 147). Whitley and Wang (2017) analyze newspaper portrayals of mental illness in Canada across a ten-year period and utilize sentiment analysis to uncover distinctions in news coverage. They found that articles increased in positive sentiment, while news categorized as stigmatizing or negatively stereotyping mental illness reduced in frequency over time. They also found differences across news genres; front

page articles and articles in broadsheets include more positive sentiment (Whitley & Wang 2017: 278).

Jaspal and Nerlich argue that using media for medical related studies is vitally important considering that media 'constitute a key source of societal information regarding scientific, technological and medical developments' (2017). Jaspal and Nerlich draw on corpus data to consider representations associated with HIV in the UK press, and use both corpus-based methods and social representation theory – namely social psychological mechanisms *anchoring and objectification* (2017: 482); they additionally pick out themes based on qualitative observations and find that media typically 'rely on metaphors and commonplace images' when discussing new medical problems (Jaspal & Nerlich 2017). Jaspal and Nerlich (2017) argue that there is evidence of competing representations of what they call the "hope" representation and "risk" representation, emphasizing the doubts and "scientific uncertainty as well as risks to human health" in press depictions and health communication (Jaspal & Nerlich 2017: 491). Additionally, health-related corpora have been found useful for considering language-specific differences related to this particular genre of writing (McEnery et al. 2006: 324-43). Finally, perhaps the most relevant and important study in this area is the work by Kretzschmar et al. on US tobacco company documents (2004). This study utilized corpus methods and principled sampling to identify linguistic and rhetorical strategies in tobacco company documents to find evidence of misinformation and demonstrates both the importance and relevance of corpus methods for health-related topics and health communication (Kretzschmar et al. 2004).

Other works highlight historical changes in news discourses over time, including representations of marginalized groups, migrants, and asylum seekers (Morley & Taylor 2012, Taylor 2014), and of Muslims and Islam (Baker 2010, Baker et al. 2013). Baker's (2010) interrogation of British press incorporates concordance lines and keyword analysis to examine representations of Islam and Muslims. Findings highlight differences between tabloids and broadsheets; tabloids primarily emphasized "British interests", while broadsheets wrote about

Muslims in a diverse array of situations, including stories from across the world (2010). McEnery and Baker's study of historical sources and discourses of prostitution (2017) makes use of the Early English Books Online (EEBO) corpus and collocations to understand meaning trends over time in Early Modern English. This study also focuses on collocations following Gabrielatos and Baker's idea of "consistent collocates" which occurred "reasonably consistently" across the duration of the corpus (2008, as quoted in McEnery & Baker 2017: 25). They further utilize other categories of collocates, including terminating, initiating, and transient to understand evidence in discourses and meaning change over the century of data studied (25-28). Partington's work on antisemitism in press accounts found distinctive trends, including increases in "a perceived resurgence of antisemitism in Western Europe" (Partington 2012: 51). Partington utilizes critical discourse analysis in addition to corpus-based methods, with frequency analysis and analysis of concordance lines (2012).

3. Methodology

This chapter incorporates a variety of methodologies from corpus linguistics. Data compilation, preparation, management, and continued analysis are all accomplished with the aid of multiple computer programs: Python (Version 3.7.4.), Unix, CWB (Corpus Workbench) and CQP (Corpus Query Processor) (Hardie 2012, Evert & Hardie 2011), R Studio (R Core Team, 2017) and several R packages, including polmineR and RcppCWB (Blätte et al. 2019). In order to avoid research errors and bias, the data are subjected to a variety of techniques (Baker 2014: 200).

3.1 Corpus Compilation and Creation

It is important when compiling a corpus to ensure creation using principled sampling methods (Kretzschmar et al. 2004)[20]. In order to construct these corpora, the University of Georgia's library databases were accessed to obtain all articles containing the keyword(s): *e-cigarette(s).* Table 1 lists all data sources, with corresponding labels for both the United States

[20] For more on sampling and representation see pages 20-23.

and United Kingdom corpora. As shown in Table 1, these data sources[21] include a variety of national and regional published news media, from many different locations and areas of the United States and the United Kingdom. They also represent a variety of different political views and affiliations. In addition to these newspapers, articles from National Public Radio (NPR) are included in the US corpus, as many Americans rely on NPR as a source of information for current news. Articles from the British Broadcast Corporation (BBC) are included in the UK corpus. Both BBC and NPR articles are included for the same date range and follow specific criteria when downloading them directly from their websites bbc.com and npr.org. Selections chosen were categorized as articles rather than interviews or transcriptions of newscasts, and they had the same requirement to contain the term *e-cigarette(s)* at least once. When downloading and organizing these data, two files were created per corpus. One file included all metadata information, organized in a spreadsheet, with separate columns for each article date, title, year, source, and geographic location. All national sources, from both the UK and US are labeled as *national*, and for all regional sources, the location was identified accordingly. For example, in the US, the *AJC* is in the location *Southeast*, and in the UK, an article from *The Herald* is labeled as *Scotland*. Updating the metadata spreadsheets was constant and consistent with every update to the plain text file for each corpus. This file contains the raw text of the articles, using a plain text editor, with article titles and body designated through the labels <*title /title*> and <*body /body*>. This was accomplished manually starting in October 2019 and continuing through March 2020, for every article to avoid repetition and to ensure that the data was included across the appropriate time spans and sources.

When both corpora were at 1 million words, a Python script (Version 3.7.4) was utilized to format the data in both corpora into CWB-encoded (Corpus Workbench format, one word per line) in order to upload it to CQP (Hardie 2012). This involved creating several scripts so that the

[21] The full tokens by article source and article year are shown in Appendices B-G, pages 167-169.

data from both metadata spreadsheets and the raw texts could be appended to each word and uploaded to CQP as structural attributes of the corpus. Python was also used to create article publication year as an additional attribute, set apart from the exact publication date.

CQP and CWB are the main tools used for corpus analysis primarily due to processing power and flexibility (Hardie 2012, Evert & Hardie 2011). These tools worked well with the data management and organization of each corpus, so that modifications and updates with the addition of new data could proceed through this established workflow. R Studio (R Core Team, 2017) and several R packages, including polmineR and RcppCWB (Blätte et al. 2019) integrate with CQP for corpus analysis (Hardie 2012), and provide flexible options for corpus methods and visualizations for analyses over time. These analyses help to determine if language patterns change across the time span of each corpus and show whether additional scientific evidence was included.

4. Discussion and Analysis: Patterns of Representation

Establishing the patterns of use and representation in the data and frequencies introduces the story of the rhetoric of e-cigarettes in US and UK press. The tokens *e-cigarette(s), vape(s|ing), health, user(s), youth,* and terms related to use of e-cigarettes are investigated via collocations[22] in order to understand and establish baseline descriptions over time. Table 5.2 presents results for the collocates for e-cigarette(s) in the complete US corpus.

Table 5.2: US E-cigarette(s) Collocates

Token	**Raw frequency**	**Log-likelihood**
use	1807	3453.2
flavored	931	2246.9

[22] Collocations are obtained in polmineR using the co-occurrences function (Blätte et al. 2019), and the default setting includes a span around the node word of five tokens to the left and right. These settings can be adjusted in the R script. The co-occurrences function computes log-likelihood statistics for each collocate based on contingency tables. Function words are not automatically excluded, but these analyses utilize a stopword regular expression (regex) from the R package tm (Feinerer & Hornik 2019) in order to remove them.

using	602	1086.9
ban	779	936.4
sale	407	816.8
used	551	757.6
users	384	617.4
makers	224	601.9
sales	514	499
products	1121	481
minors	303	463.7

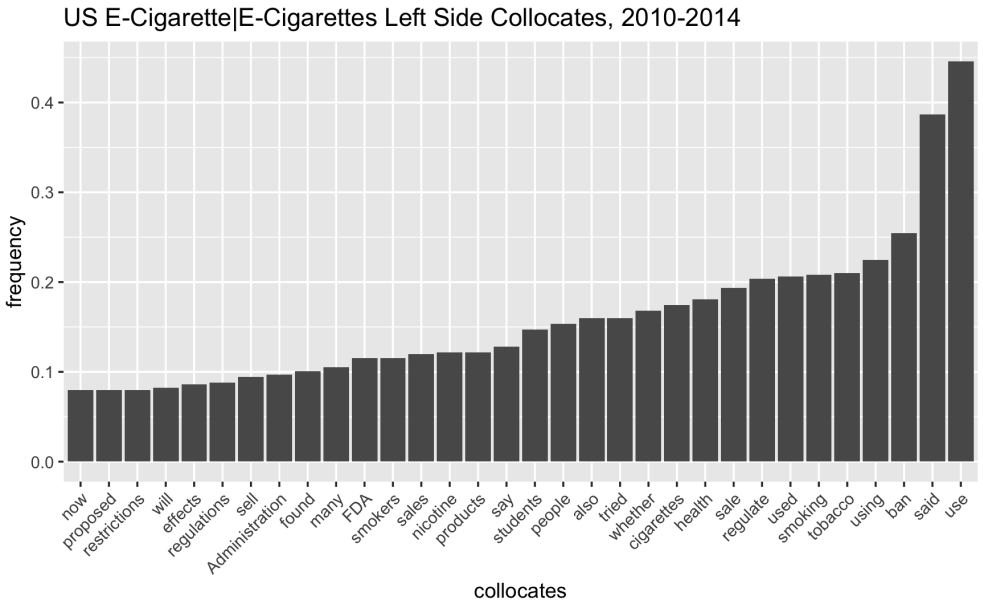
The top collocates in the US corpus reflect an overall concern with e-cigarette use and users. A link is established between discussion of users and minors, with such collocates as *students* and *youth.* There is also global discussion of trends with e-cigarettes: *banning, selling, regulate,* and *tried.* It is notable that the top collocates reflect additional concerns with industry, regulation, and the market through tokens such as *products, companies, manufacturer, market, maker(s), regulate,* and *sale.* Table 5.3 displays the results for the top collocates of the same group, *e-cigarette(s)* for the UK.

Table 5.3: UK E-cigarette(s) Collocates

Token	**Raw frequency**	**Log-likelihood**
use	790	1549.7
using	375	687.9
users	275	512
flavoured	164	444.6
ban	313	406.3
used	269	311.9
sale	157	286
that	358	3.16
vapour	322	70.37
market	308	425.11

The UK results center primarily around e-cigarette use and users, through the top collocates *use, using,* and *users.* Additional noteworthy top collocates include *ban, banned, sale, market, flavoured,* and *advertising,* as well as *evidence* and *safe.* Consideration of these topics in context reveals ongoing concern regarding user safety and whether e-cigarettes are actually safer than other tobacco products on the market. This comparison can also be made over time and more specifically paired down into left and right side collocates[23]: the collocates that most frequently modify the terms *e-cigarette* and *e-cigarettes* and those tokens which are most frequently modified by the latter terms. These were accomplished by dividing each corpus into subcorpora by year groupings, so that each group had a similar number of tokens[24] for a better sample and comparison. The resulting groups for the US are 2010-2014, 2015-2018, and 2019-2020. The groups for the UK are 2010-2014, 2015-2017, and 2018-2020.

Figure 5.1



[23]In order to obtain collocates for right and left side only, a simple adjustment of the polmineR script is required, where right or left is set to zero (Blätte et al. 2019).
[24] The numbers are shown in Appendices F-G on page 169.

Collocates were obtained for each time period in these subcorpora so that the values could also be normalized per 1,000 tokens, for comparing use over time. In the first time group for the US, *use, said, ban,* and *using* are the four most frequent collocates modifying e-cigarette or e-cigarettes. Other collocates relate to business, regulation, and the government: *regulate, regulations, sale, products, sales, FDA, Administrations,* and *restrictions.* In addition, collocates related to people and users of e-cigarettes are present: *health, smokers, people,* and *students.* There are also a great many verbs in this group, potentially showing that e-cigarettes are passive agents or receivers of action in the press at this point in time in the US.

Figure 5.2

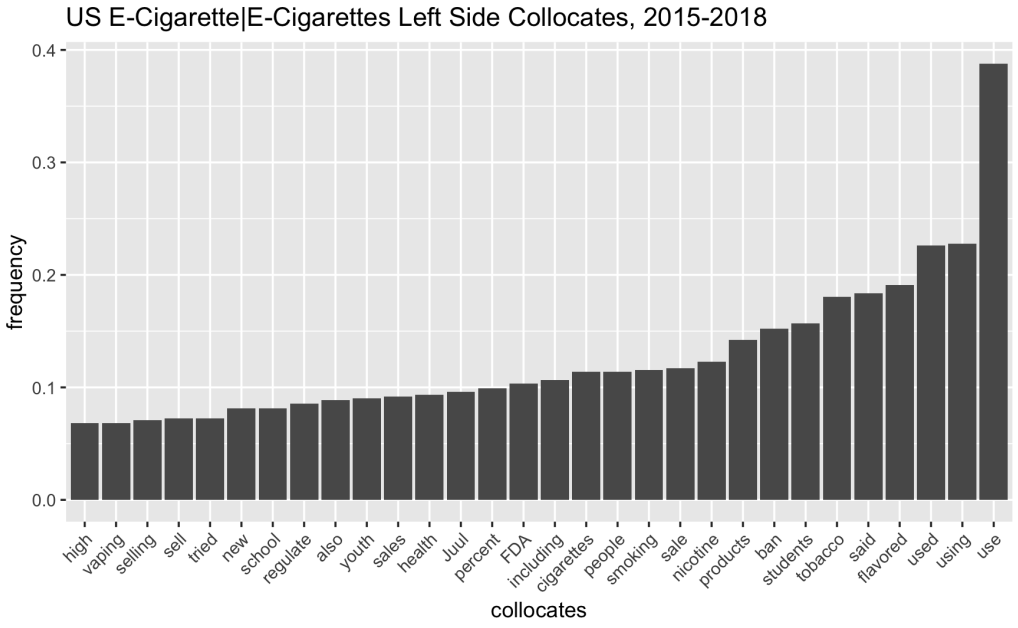
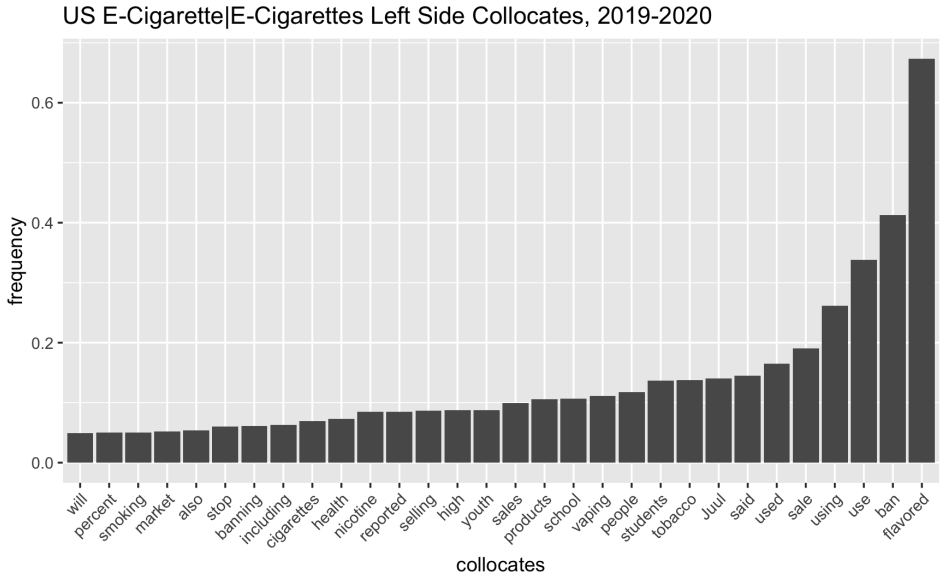


Figure 5.2 displays results for the next time group in the US, 2015-2018. From 2010-2018, *use* remains a stable top collocate. A new top collocate, *flavored,* highlights changes in technology related to these devices. An additional development in these frequency profiles is that *ban* moves further down on the distribution, as compared to 2010-14. Groups of terms related to users are present, including *people* and *health,* which remain stable collocates. *Students* occurs with greater frequency than in 2010-14, and *youth* and *school* are newly introduced. Terms

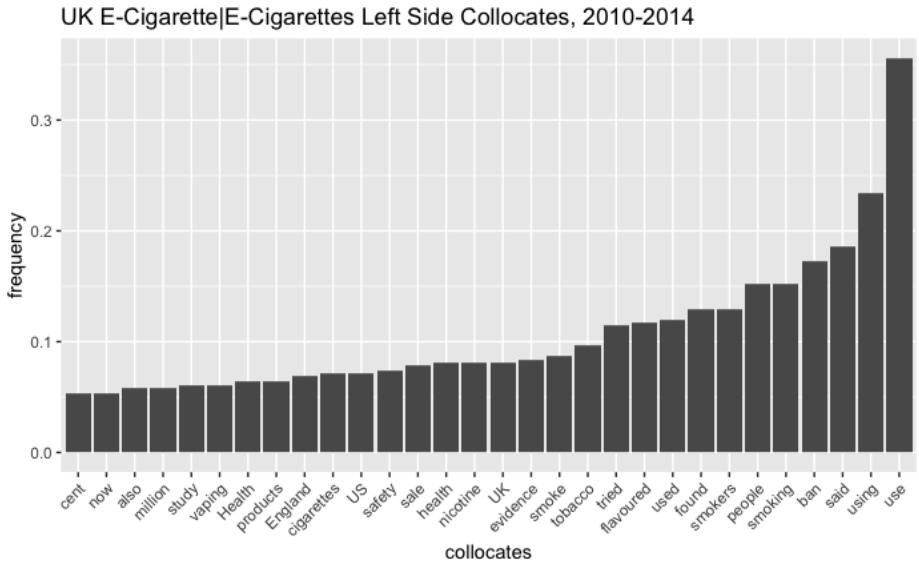
related to these devices as *products* rise in frequency: *sale, products, Juul, sales, regulate, selling, sell, sales.*

Figure 5.3



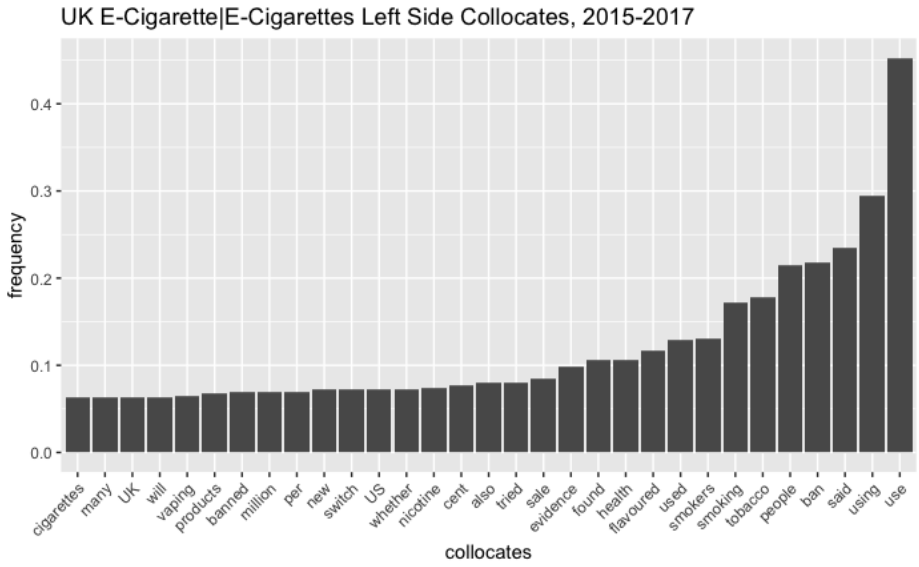
The left side collocational profile for 2019-20 reflects interesting changes in the discussions surrounding e-cigarettes. *Flavored* and *ban* surpass *use* as the top two collocates. *Juul, sale, students, vaping,* and *youth* also rise in frequency, while *products, people,* and *sales* remain stable. The collocates *smoking, nicotine,* and *health* are lower in frequency. There are also fewer verbs in this group, indicating a movement away from framing e-cigarettes as passive agents. While more collocates are stable early on in the corpus, there are greater changes beginning in 2015 and into the final time group. These changes could be due to mounting evidence or simply the rise in use of e-cigarettes in the US, in addition to more regulations being imposed on e-cigarettes worldwide. Interestingly, these collocates are not overwhelmingly negative or positive. Next, the results for collocates modifying e-cigarette(s) are displayed over time for the UK.

Figure 5.4



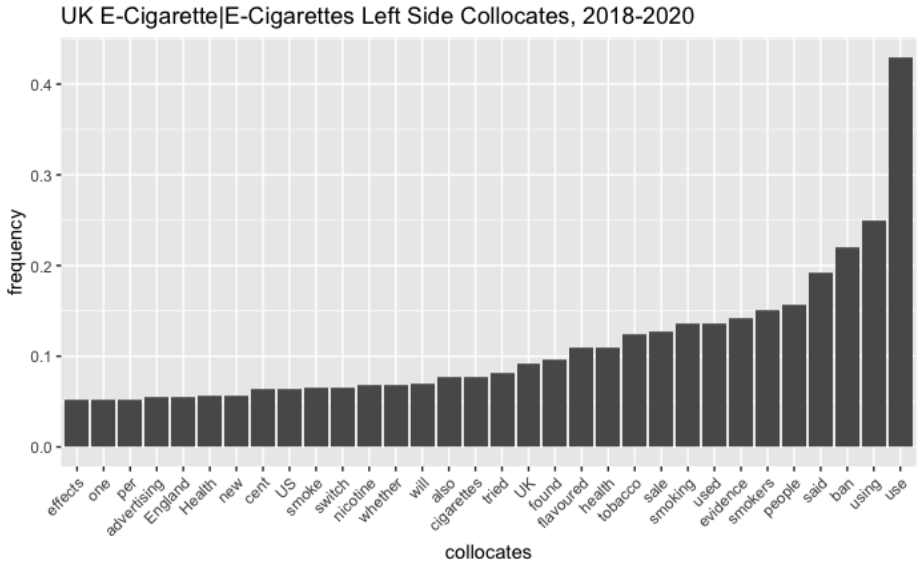
Moving across the Atlantic, the top collocates in the UK in 2010-2014 are *use, using, said, ban,* and *smoking.* Collocates related to people include: *smokers, people,* and *health,* which are of higher frequency than tokens related to business, government, and regulation: *advertising, sale, selling, market, regulate, rules,* and *companies.* Other remarkable collocates are *evidence, England, US, study,* and *found,* none of which are present in US press.

Figure 5.5



In 2015-17, *use* remains a stable collocates, as well as *people, ban, smoking,* and *smokers.* Overall there is great similarity in the frequency distributions since 2010-14, and *health* and *evidence* are notable stable collocates since that time. Terms including *found, evidence,* and *health* relate to research in the UK press. *Switch, new, whether,* and *UK* are collocates that are newly introduced in this time frame.

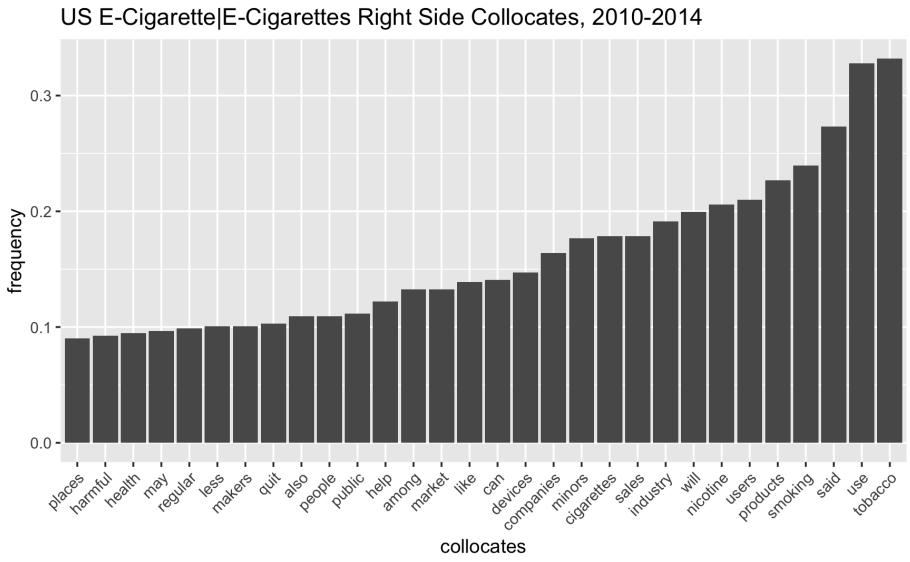
Figure 5.6



Use and *using* remain the stable top two collocates throughout time in the UK. *Ban* rises in the frequency profile in 2018, and tokens related to smoke: *tobacco, smokers, smoking* remain stable throughout 2010-2020 in the UK. *One, linked,* and *effects* are new collocates introduced in 2018. There are notable differences found in these collocates between the US and the UK; *evidence, research, found,* and *study* could indicate differences in press reporting. If the UK reports are citing outside research, and this is not present in the US, then these reports are providing a more comprehensive outlook on the issue. There are also similarities present in the collocates, including a lack of overwhelmingly positive or negative collocates, government and regulation related entities, and terms related to users. In the UK, users are identified as *smokers* rather than as *youth.* Similar tokens that are stable throughout both include *use, ban,* and terms related to regulation.

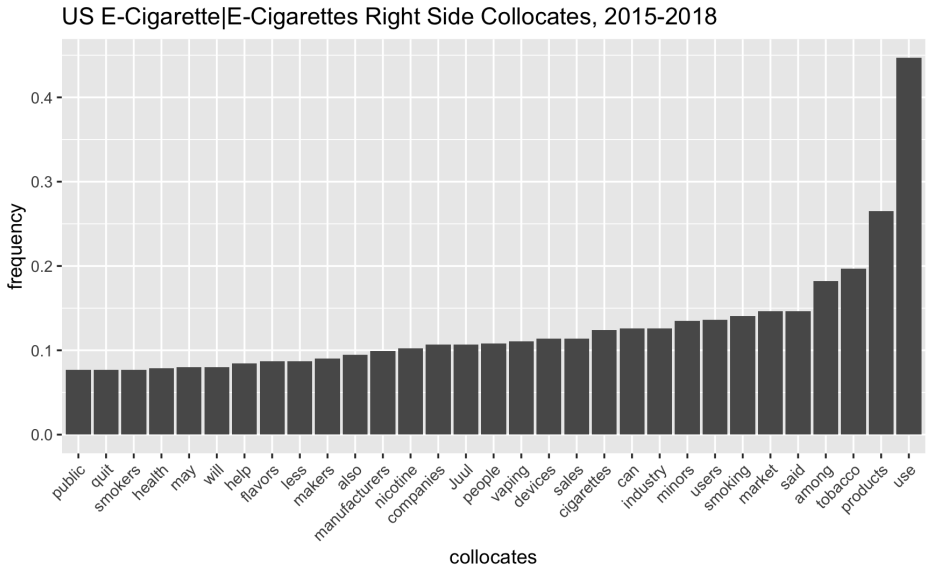
Right side collocates for both the UK and the US show differences over time from both corpora and from the left side collocates. These are the terms that are being modified by *e-cigarette* and *e-cigarettes.* Figures 5.7-5.9 shows these terms over time for the US, and figures 5.10-5.12 displays results for the UK.

Figure 5.7



The right side top collocates are stable throughout 2010-2020 in the US press: *tobacco, use* and *products.* In 2010-14, collocates related to traditional tobacco products are present: *tobacco, smoking, nicotine,* and *cigarettes.* Another group of collocates is related to business: *market, products, companies, markers, sales,* and *industry.* Additional tokens of note are *harmful, health, people, public, users,* and *minors.*

Figure 5.8



Many of the collocates introduced in 2010 remain present in 2015-18, but others still are introduced including *Juul, vaping, manufacturers, help, flavors,* and *may.* The terms *may* and *help* promote the unclear version of events in US press during this time. *Juul* rises in frequency by 2019-20. The tokens *young, ban, pods,* and *Labs* are also introduced in this time. *Labs* is a proper term connected with *Juul Labs.* These collocates are not surprising given the changes in e-cigarette use and overwhelming concerns with youth. Yet, tokens related to the industry and markets are still stable throughout, showing this is a major feature and characteristic of US press.

Figure 5.9

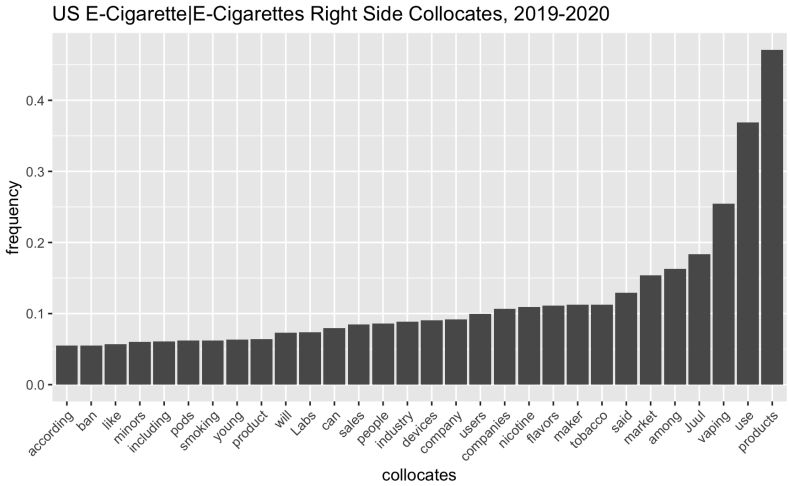
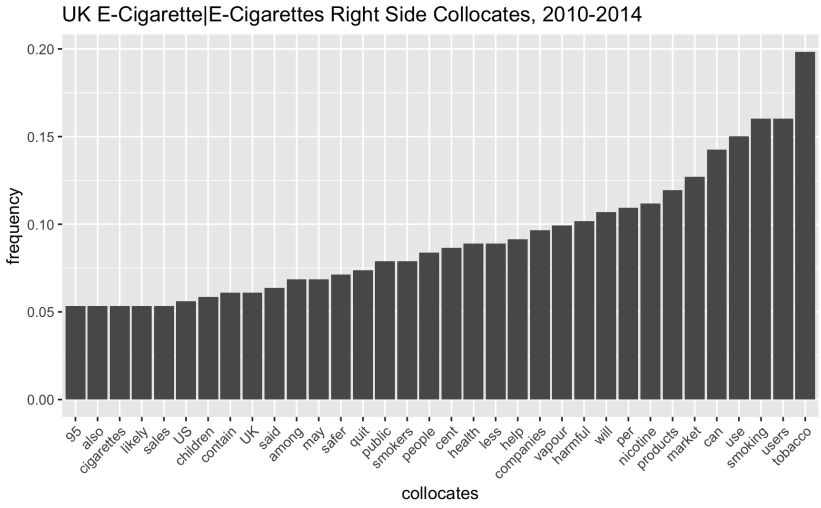


Figure 5.10



The collocational profile in the UK is both different from the US and changes over time. In 2010-14 the top tokens are *tobacco, users,* and *smoking.* In 2015, *products* is of much lower frequency and *use, users,* and *tobacco* are the top three collocates, and *vaping* and *vapour* are introduced. In 2018-20, *use* remains stable as the top collocate, followed by *smoking* and *tobacco*. Interestingly, a quite mixed group of collocates are present also: *quit, safer, harmful, banned,* and *help. Harmful* and *help* decline in frequency over time. In 2018 new collocates are introduced: *banned, industry,* and *among.*

Figure 5.11

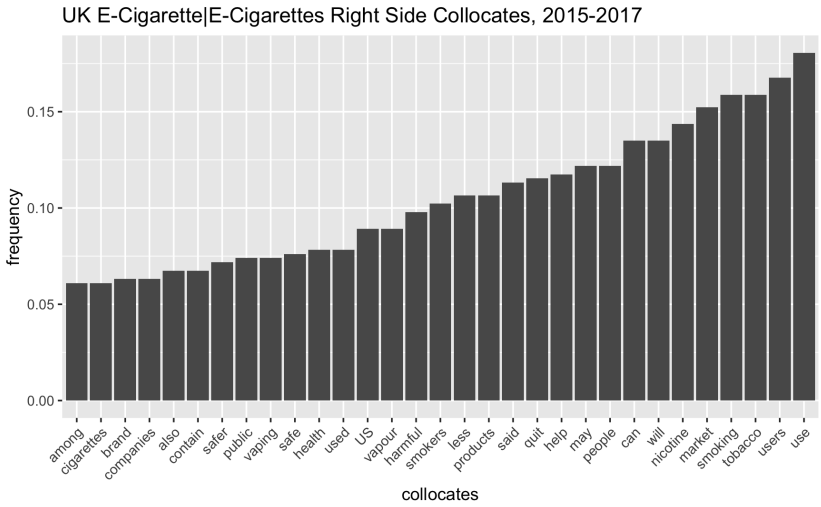
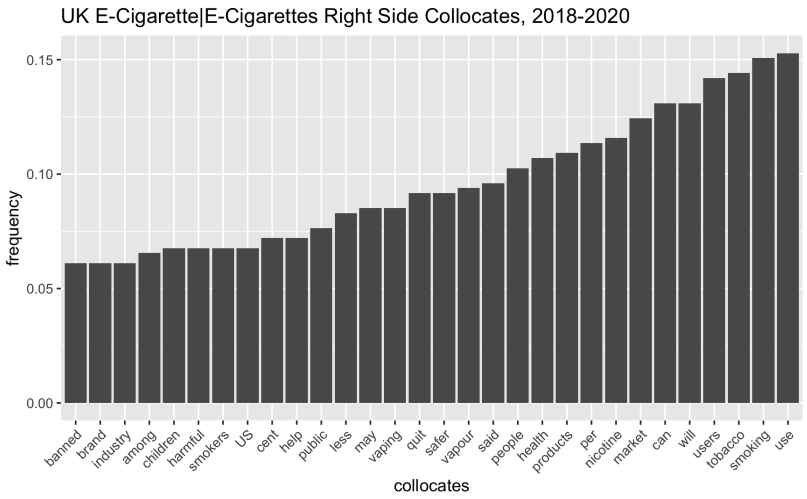


Figure 5.12



In order to further expose the rhetoric and patterns surrounding e-cigarettes, the collocates for terms *vape, vapes,* and *vaping* are presented in the next tables.

Table 5.4 Vape(s|ing) US

Token	**Raw Frequency**	**Loglikelihood**
products	1574	2484
shops	407	1626.6
THC	366	1088.6
linked	200	614.7
youth	350	611
pens	132	588.9
flavored	371	509.7
teen	180	473
teenage	122	401
epidemic	170	393.3
devices	353	380.5
shop	131	368.5

Table 5.5 Vape(s|ing) UK

Token	**Raw Frequency**	**Loglikelihood**
linked	102	294.4
products	224	220.5
switch	71	179.88
is	654	163.2
less	117	154.44
pen	31	132.25
THC	38	114.08
safer	72	105.44
teenagers	62	105.16
people	226	94.16
liquids	35	92.35
evidence	91	90.62

When comparing these results, it appears as though the US corpus is particularly focused on business aspects of vaping, and vaping pens as a product. This is evident in the terms: *shops, shop, stores, devices,* and *products.* The UK corpus as a whole does appear to consider these as a product but is more concerned with the effects of vaping on people as individuals (shown in table 5.5). Examples of tokens related to this are *safer, people, teenagers,* and *evidence.* Terms referencing e-cigarettes as a product in both corpora are *pen(s), liquids, THC,* and *products.* It is also evident that linguistic patterns surrounding terms *vape(s|ing)* and *e-cigarette(s)* have their own extensions and patterns that are unique, with a tendency to be connected to this technology as a product in both the US and the UK, rather than in connection with research or scientific evidence, which is only present in the UK. It appears that the US news is promoting a tendency to focus on the business and governmental side of vaping, and secondarily on users as people themselves, while in the UK, it seems as though the press is focused on users as individuals, people, with secondary focuses on regulation, industry, and research. The next group of terms highlight changing descriptions of e-cigarette users themselves, and terms of interest for public health.

Table 5.6: US *Health* Collocates

Token	**Raw frequency**	**Loglikelihood**
public	1579	7515
officials	707	2726.7
risks	307	1396
care	252	1180
effects	276	1136
experts	225	923.3
department	147	646.9
advocates	170	624.2
crisis	131	508.8
departments	75	473.9
groups	140	424.6
Public	136	423

The top collocates for the US with the term *health* include a variety of tokens, although most are related to government, industry, and regulation. Additional tokens that are of high frequency include *risks, care,* and *effects*. It might be expected that additional tokens related to youth or underage users would be present, but it appears as though these are the main concerns of the US news press as a whole. The entire list for collocates related to health is quite interesting, however, and terms that are lower in frequency include some proper nouns related to places (*Michigan, Wisconsin, Boston)* where health is of concern. There are also a large number of verbs including *poses*, *threats*, *carry*, *recommend*, *assess*, *worry*.

Table 5.7: UK *Health* Collocates

Token	**Raw frequency**	**Loglikelihood**
public	611	2853.3
risks	153	628.8
officials	116	588.3
experts	150	531.5
mental	73	441.5
warnings	83	422
professionals	64	361.9
effects	113	340.6
benefits	77	335.6
minister	66	315.9
professor	67	293.8
impact	68	231.2
long-term	79	227.4

Terms in the UK reflect differences in health and governmental systems between the US and the UK: *minister, director, chiefs.* Similar terms are also present, and it is not surprising that *public* and *risks* are both top collocates in the US and in the UK. Table 7 shows the results for collocates related to health in the UK. Tokens in both corpora reflect larger concerns with public health, industry, and government, rather than viewing or describing health as something which also applies on multiple levels of scale, down to individual decisions and choices. The lack of mention of local governments and individual levels is an area that should be of concern to national and regional press as well as how public health decisions and outcomes affect individuals and groups locally.

Next, collocational analysis results are discussed for an important group of terms related to e-cigarettes: use-related terms and collocates for *user(s)* in the US and the UK. Tables 5.8 and 5.9 show the results for the collocations query in each corpus, for: *use*, *uses*, *used*, and *using*. Tables 5.10 and 5.11 display results for collocates of *user* and *users* to uncover the patterns of language use in the press with regard to who uses *e-cigarettes* and how they are characterized as individuals and as a group.

Table 5.8: US Use-related collocations

Token	**Raw Frequency**	**Loglikelihood**
e-cigarettes	1782	2938.3
e-cigarette	1104	1894.1
among	564	1253.7
students	500	1001.3
youth	410	942.8
products	915	837.7
school	421	704.9
tobacco	827	704.9
high	388	699.4
who	743	659.9
underage	178	526
reported	283	500
teens	212	367

Table 5.9: UK Use-related collocations

Token	**Raw frequency**	**Loglikelihood**
e-cigarettes	1071	1963.5
people	464	599.6
them	330	454
e-cigarette	345	426.2
devices	171	283.6
who	339	288.8
among	127	250.9
million	129	226.7
their	298	221.7
electronic	131	189.5
quit	159	187.9
smokers	214	171.1

Tables 5.8 and 5.9 show the results for the US and UK respectively for collocates of these terms. Top tokens are similar for both *e-cigarette(s).* The US collocational profile includes far more tokens related to youth: *students, youth, school,* and *underage.* The UK includes references to people in general as users: *people, them,* and *who.* Additional top collocates include *million*, *among*, *their*, *smokers*, *quit*, *young*, and *electronic*, as well as *T/teenagers*. The terms *product* in the US corpus, and *devices* in the UK corpus also reveal a tendency to characterize e-cigarettes as passive products of use in general. Users are characterized in different ways shown in their

respective collocations, in tables 5.10 and 5.11 below. Interestingly, similar characterizations of users include reference to actions, with collocates like *inhale* and references to products in use: *e-cigarette(s), nicotine, non-nicotine,* and *tobacco.* Also, the US includes *underage* as a top collocate, while the UK characterizes uses are people or smokers: *they, their,* and *smokers.*

Table 5.10: US User(s)

Token	**Raw frequency**	**Loglikelihood**
inhale	110	757.5
e-cigarette	303	743.2
vapor	98	308.9
underage	61	247
inhales	30	241
nicotine	146	186.6
Non-nicotine	20	169.4
expose	18	125.5
e-cig	33	124.8
current	38	118.8
dual	14	111.4

Table 5.11: UK User(s)

Token	**Raw frequency**	**Loglikelihood**
e-cigarette	205	615.7
nicotine	64	66.6
e-cigarettes	63	8.9
have	59	3.7
were	56	40.9
inhale	54	355.8
they	52	8
their	46	17.7
more	45	16
smokers	42	29.5
tobacco	40	4.7
million	37	82.3

Further collocational analysis is warranted because concerns over youth use of e-cigarettes are involved in a large amount of scientific literature. As discussed above, *youth* and tokens related to youth are present in those collocations. Tables 5.12 and 5.13 display results of *youth* collocations in the US and the UK.

Table 5.12: Youth in the US

Token	**Raw Frequency**	**Loglikelihood**
Survey	151	1287.5
use	357	969.4
National	138	670
vaping	337	662.3
epidemic	132	342.67
Tobacco	108	248
among	41	239.7

usage	58	232.4
our	45	162.5
access	47	158
appeal	34	144.4
reduce	48	155.7
combat	34	149.22

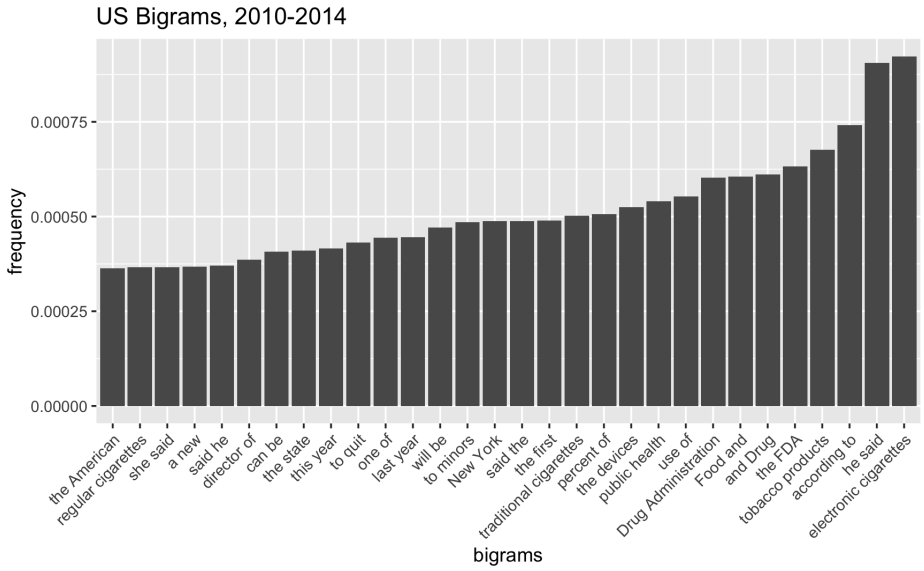
Table 5.13: Youth in the UK

Token	**Raw Frequency**	**Loglikelihood**
use	33	95.3
culture	11	80.3
vaping	30	57.5
epidemic	9	51.9
affected	7	41
Survey	5	36.8
crisis	6	35.15
appeal	7	35.12
linking	4	32.6
our	13	22
today	8	32.6
address	5	29.7

Shown above are the results for collocates for *youth* in the US and the UK, and one of the top tokens for both is *use.* Notable tokens in the US include *appeal, reduce, combat, use, usage,* and *epidemic,* and even though the tokens are of much higher frequency in the US than the UK, there is still interesting data in the UK corpus. Notable tokens in the UK are *culture, crisis, epidemic,* and *affected.* Interestingly, *New York* is also included as a collocate in the UK corpus, as well as the proper noun, *America*. In order to further understand the changes in rhetoric over time, bigrams[25] are calculated for each subcorpus using the year attributes[26] in the queries for each corpus.

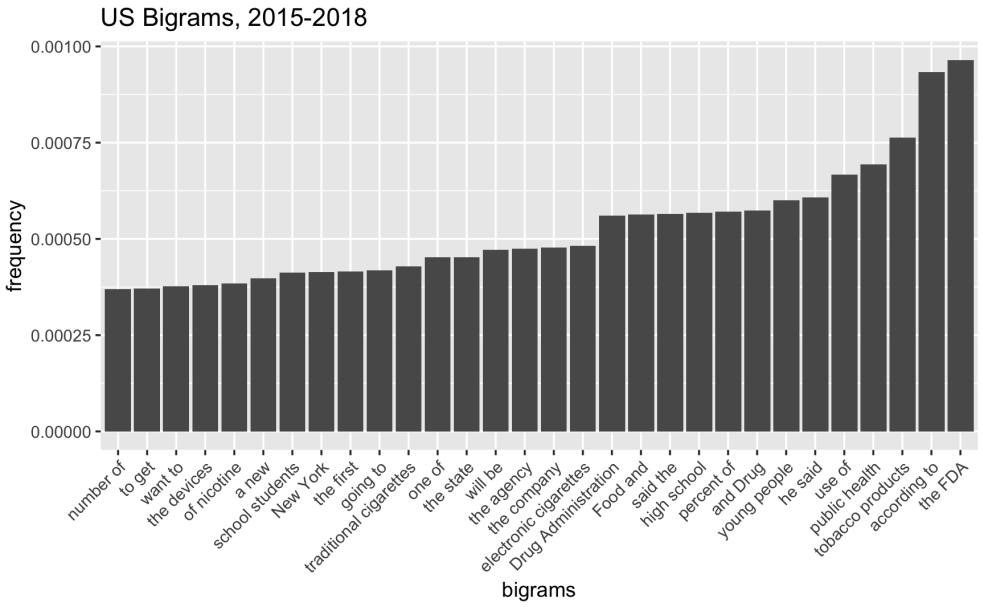
[25] In order to obtain bigrams over time for each term, the ngram function in polmineR is used and then bigrams are ranked by raw frequency and adjusted per 1,000 tokens. The bigram function in polmineR does not automatically include statistical measures but instead outputs the raw counts for each bigram. Stopwords are also employed so that only one function word can be present in a bigram pair.
[26] Metadata, such as article date of publication, article year, and source, are all stored in the corpus as s_attributes with CQP. The attributes are then easily accessible via queries in polmineR.

Figure 5.13



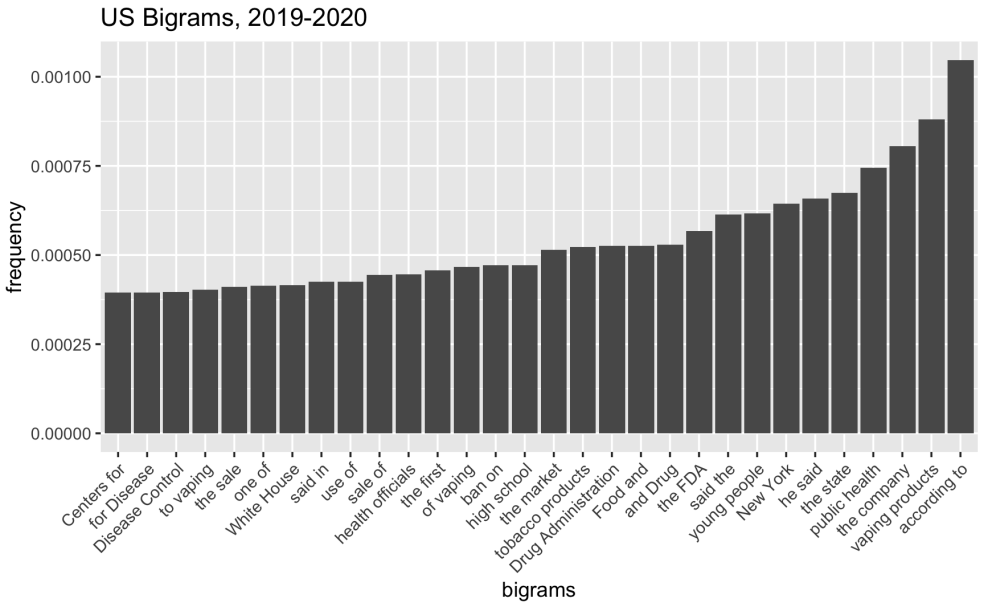
The top bigrams from the years 2010 to 2020, reflect market and overall concerns in US news over time. Some of the bigrams are stable throughout each time frame, and others are not. In 2010-14 in the US, top bigrams include *he said, tobacco products,* and *the FDA.* The figure above shows content bigrams with consistent linguistic patterns related to regulation and e-cigarettes as products. There are also associations with tobacco, beginning in the 2010-2014 group which continue to demonstrate stability throughout the study period. Some of these bigrams include *tobacco products, regular cigarettes, to quit* and *the devices. Public health* is another notable bigram, which is stable throughout the US corpus, though it is of much lower frequency in this first year grouping. In 2015-18, the top bigrams are *the FDA, he said,* and *public health.* Additional bigrams introduced by the press at this time are *high school, young people, school students, the agency,* and *the state.* These point to a preoccupation and concern with student use of e-cigarettes.

Figure 5.14



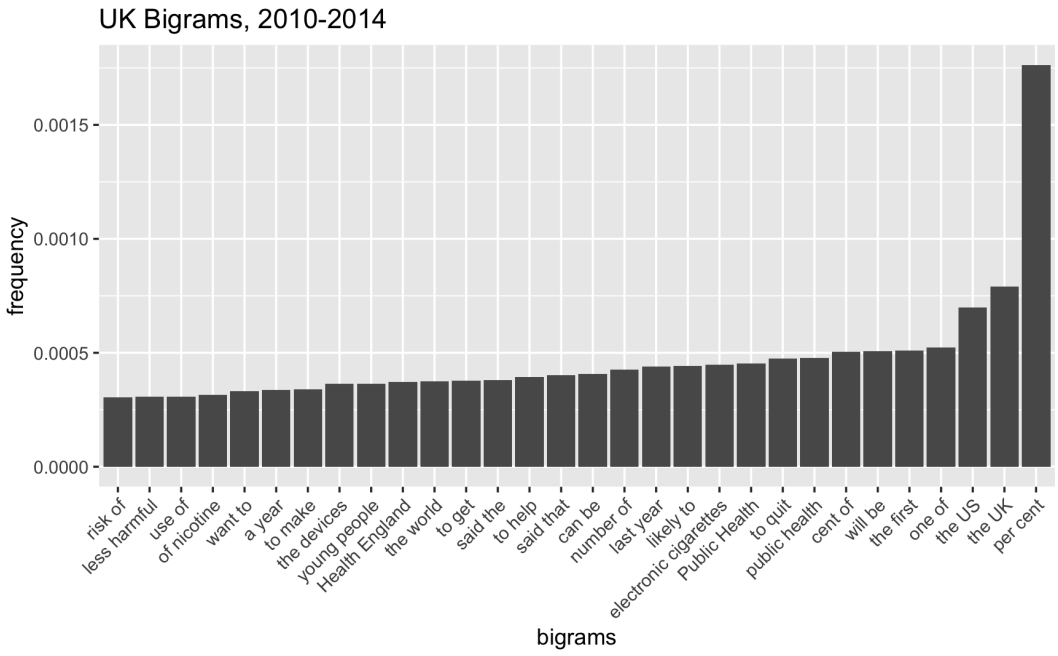
Electronic cigarettes and *tobacco products* are the most frequent content word bigrams up until 2018. Throughout the duration of the corpus, a variety of bigrams are used in reference to government, business, and regulation, which reaches a peak in 2015-18 with *the FDA, Drug Administration, the company, the agency,* and *the state*. Tokens referencing the CDC do not occur until 2019: *Centers for, for Disease, Disease Control.* New bigrams introduced in 2019 include: *vaping products, ban on, White House, the sale,* and *to vaping.*

Figure 5.15



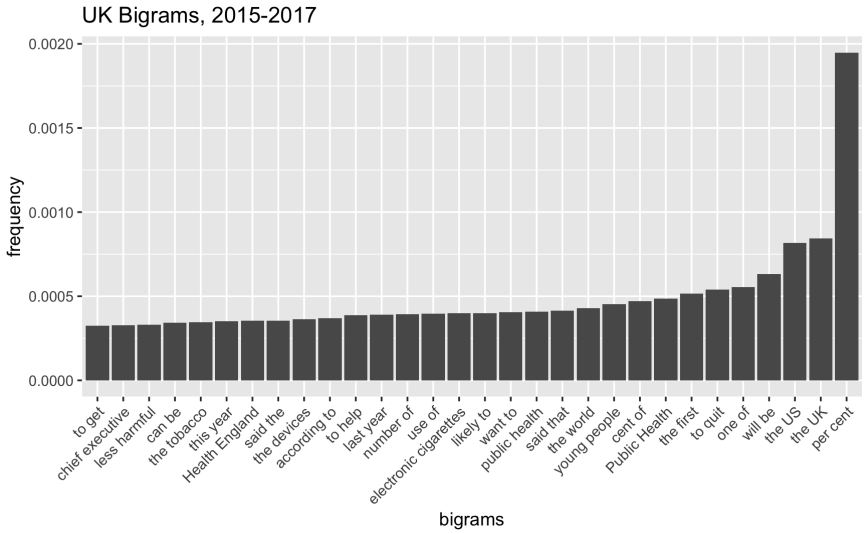
In 2019-20, top bigrams are *according to, vaping products, the company,* and *public health. According to* is a bigram pattern that is also characteristic of academic writing (Biber, as quoted in Greaves and Warren 2010: 295). The US press bigrams relate to youth and governmental regulation, which is in keeping with rhetorical positioning present throughout the US corpus. Results for bigram analysis over time in UK press are shown below, reflecting interesting differences between the US and UK patterns. Figure 5.16 shows the content bigrams for the UK in 2010-2014.

Figure 5.16



Notable top bigrams in this group are *the UK, electronic cigarettes, public health,* and *to quit.* In line with the collocates that are present in UK press, these bigrams present characteristics of rhetoric that are not present in the US, as well as references to other parts of the world: *the world, the US. Health England* and *Public Health* are bigrams that reference *Public Health England,* a governmental entity. Different groups of content bigrams relate to tobacco products, like *the devices* and *of nicotine*. Other bigrams present this news topic in a more positive light- *the first, to help,* and *to quit.* Stable bigrams are present in both corpora from the first year grouping.

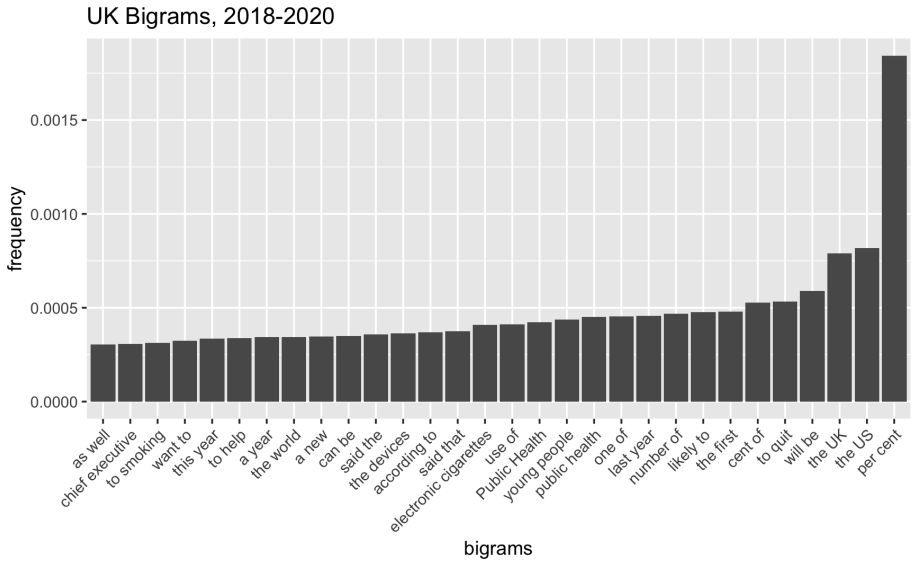
Figure 5.17



The UK remains the second most frequent bigram through 2017, and *to help, public health, to quit, less harmful,* and *the devices* are stable bigrams through this time. Several new bigrams occur in 2015: *chief executive, the world,* and *according to* . This brings up an additional characteristic of UK press that is not prevalent in the US: higher mentions of research, evidence, and outside sources.

Public health and *to quit* remain stable throughout the UK press over time, but the bigram *use of* rises in frequency through 2020. *The US* surpasses *the UK* as the second most frequent bigram in 2018. *Young people* also rises in frequency starting in 2018, while *the world, according to,* and *to help* are lower on the frequency profile at this time. Bigrams using the term *vaping* occur with much greater frequency in US press. Other bigrams that are newly introduced in 2018 are *likely to, want to, less harmful,* and *the tobacco.* Intriguingly, there are no direct references to people in the UK press in bigram patterns other than *chief executive* in 2015-20 and *young people* in 2010-20.

Figure 5.18



These results do show varying concerns with relevance to the US and the UK press. In the above analysis, there are a variety of methods presented that show different types of linguistic patterns over time. The collocations surrounding terms *e-cigarette(s), user(s), youth, vaping,* show the specific contexts in which these terms were used in both the US and the UK and provide baseline descriptions of each. Their use and frequency patterns change over time, underscoring the fact that linguistic patterns will change as technology changes. Prior to the time of these corpora, e-cigarettes existed; they were not widely used or distributed, and even today, the health effects of e-cigarettes are not entirely clear (cassaa.org). The implementation of collocational analysis provided a unique snapshot of the linguistic data. Finally, results of bigram analysis over time are presented using the actual frequencies of bigrams. Use of this method underscores larger patterns present in both corpora and stable patterns over time. In fact, there were a surprising number of stable bigram patterns for each corpus individually and for both corpora together. This effort highlights additional interesting research questions that warrant further investigation.

What emerges from the combination of methods above are characteristics of an important and especially relevant topic of news today. Both US and UK press change their stories over time,

and they also differ in key ways. Similarities involve concern with the governmental systems in each location, with reference to business and markets being emphasized more consistently in US press; another important similarity is that both US and UK press are not overwhelmingly positive or negative in characterizations of e-cigarettes or in language patterns over time. Key differences include the UK's reference to scientific research and evidence, with users typically being characterized as people in general, while in the US they are increasingly concerned with youth as users over time.

5. Conclusion

This chapter highlights the story of representations of e-cigarettes in US and UK press over time. The rhetoric surrounding e-cigarettes in the US and the UK has certainly changed since 2010, but a key similarity is the vague discussions surrounding the health effects of this technology. E-cigarettes have held a contentious place in society since their introduction[27]. In order to garner a more complete understanding of this rhetoric, the incorporation of collocational and n-gram analysis provides different perspectives of the linguistic items in the corpus and different levels of analysis. Each method offers a different vantage point of the data: the use of n-grams shows the most frequent patterns and changing patterns over time, and collocational analysis shows how certain topics and tokens are described and patterned over time; for instance, the terms *health* and *youth* do not have the same frequency distributions at any point in time.

Overall, the US and the UK press focus on different concerns regarding e-cigarettes, though similarities are present such as the references to government and regulations, as well as neither overtly positive or negative language about e-cigarettes. A preoccupation with business, products, and markets, as well as users being underage is apparent in the US data. The US corpus

[27] E-cigarettes were first marketed in China in 2003 and appeared on the UK market in approximately 2007 (Royal College of Physicians Report 2016). Currently, e-cigarettes are incredibly popular with young adults, and the use of e-cigarettes is higher among high school students than adults in the US (surgeongeneral.gov). In addition, a recent study from Ash YouGov Smokefree Youth Great Britain found that since 2015, there has been an increase in youth use of e-cigarettes (June 2019).

is also centrally focused on US internal affairs; in contrast, the UK includes reference to other parts of the world, specifically to the US, and more reference to users as people in general, although there is some reference to young adults. The UK press is also more attentive to research and citing outside evidence. What happened with the data over time shows differences in each place, partly due to technology but also changes in press rhetoric.

This research is ongoing and demonstrates the utility of corpus linguistics when considering a relevant and pressing issue in today's society, like issues of public health communication. Linguistic investigation of popular news accounts helps with understandings of public perception and changes over time, as well as geopolitical differences in government regulation and use. Relevant research should include further investigation of linguistic indicators in press accounts – especially when there is not a clear definition of health effects of a new technology or a disease. Do these indicators differ by place or depending on governmental infrastructures? Between the US and the UK, major differences are present in health care systems and in governmental regulation of items that include e-cigarettes and other tobacco products. Additional research questions also include whether regional and national press differ significantly, and whether further lexicogrammatical differences exist between US and UK press discussions of e-cigarettes.

CHAPTER 6

WASH YOUR HANDS: CDC, NHS, and WHO TWEETS IN THE #COVID19 PANDEMIC

"Historical events and social factors change the messages spread by public health officials… both the health message and the pictures used to illustrate that message need to be carefully chosen." (WHO 2009)

1. Introduction

Throughout the year 2020, three health organizations held major roles in managing and responding to the coronavirus pandemic. The Centers for Disease Control (CDC), the World Health Organization (WHO), and the National Health Service (NHS) operated in each respective area of governance: the US, the UN, and the UK, with worldwide influence and impact. These organizations were at the forefront of countless important decisions and crucial research on the virus. In the midst of this vast effort and coordinated work, each organization also utilized social media, including Twitter and others, to communicate with individuals and the public. On Twitter, the official accounts of the CDC, WHO[28], and NHS often tweeted daily, sometimes more, tailoring communication to promote efforts concerning health and the impact of COVID-19. In the following chapter, this important dataset is analyzed.

This work tracks public health messaging and evidence of stability and change in corpora of CDC, WHO, and NHS official account tweets throughout 2020. Using corpus-based methods, similarities and differences are identified between tweets by each organization. Corpus analysis

[28] The World Health Organization accounts tweeted in many languages, including English. The chapter that follows focuses only on the English communication at present.

offers a particularly productive viewpoint by taking all communication from each source in the aggregate (as corpora) to note what topics and linguistic patterns were made consistently and most often over time. Larger macro-level and micro-level discourses and linguistic patterns are revealed, with implications for further areas of study. This chapter provides a window into the ways that governmental units communicate to the wider public via Twitter concerning public health crises more broadly and specifically regarding the COVID-19 pandemic.

In the quotation that begins this chapter, the WHO discusses the crucial impact that events have on public health messaging; all health messaging involves subjective choices by the speaker/writer, and is influenced by the larger spheres of political, scientific, cultural, and economic circumstances (Larson 2020, Kretzschmar et al. 2004, Monaco 2016, Collins & Nerlich 2017). The connection between the linguistic patterns and events is not simple, and language plays a "key role in the portrayal of risk and science" (Tang & Rundblad 2017: 667-8). Changes in scientific writing and communication reflect changes in scientific thought (Monaco 2016: 500). Since their establishment, health messaging and communication to individuals and the wider public have been an important facet of these organizations' larger goals to promote health for individuals across the world.

2. Background and Related Literature

Twitter is a unique and interesting social media source for this work, as it allows users to post short messages of 140 characters or less; it first launched in 2006. Twitter was selected as the social media choice for creating these corpora in part because it is the "most commonly used platform by public health agencies" on social media (Park et al. 2016: 188). Park et al. argue that public health groups and agencies use Twitter "primarily for the one-way sharing of information without considering audiences' needs and preferences" (2016: 190). The CDC and WHO "actively maintain a social media presence on sites like Twitter" (Yun et al. 2016: 67); the NHS has also maintained an active social media presence on Twitter (Whitelaw 2011).

Each organization was founded after the recent dénouement of World War Two with the unique backdrop of developing views on science and health. The UK and US were both under economic and political strain (Etheridge 1992, Jacobs 1993: 1-4), and worldwide, economic, political, and other pressures and problems persisted (Tassava 2008). Nonetheless, each entity was founded with overall principles supporting freedom and education and larger hopes and dreams "of defeating the microbes of infectious disease" (Jacobs 1993, Honigsbaum 2019: 7, Cueto et al. 2019: 1, 10-12) and provides a particularly interesting comparison. They also have key differences, both initially when founded and today, providing further areas of comparison.

The CDC, known as the "Nation's Prevention Agency" opened in Atlanta, Georgia, on 1 July 1946. The acronym CDC at the time stood for The Communicable Disease Center until 1992, when the CDC made an official name change to the Centers for Disease Control and Prevention as part of the Preventive Health Amendments (MMRW 1992, Etheridge 1992: xvi). The goals of the CDC were initially to eradicate malaria and similar threats in the US (Etheridge 1992: xv) but today include myriad objectives like "detecting and responding to health threats" and "taking the health pulse of the nation 24/7" (CDC 2021). The CDC's mission also includes "protecting Americans from health and security threats", "detecting and responding to health threats", "tackling health problems that cause death and disability for Americans", and nurturing public health (CDC 2021).

The WHO, like the NHS and CDC, was formed right after World War Two and was established as a special agency within the United Nations "to coordinate health affairs" (WHO Press 2007). Initially, priorities were focused on attacking malaria and tuberculosis and on improving sanitation along with promoting "the highest possible level of health by all peoples" (Cueto et al. 2019: 1, 10-12). The WHO's administrative headquarters is located in Geneva, Switzerland, and is governed by a board of health officials to coordinate efforts in each area of work. Today's goals of the WHO include "producing health guidelines and standards", "helping

countries to address public health issues", and "supporting and promoting health research" (WHO Press 2007).

Following two decades of preparation, the NHS was officially established on 5 July 1948, by the newly elected Labour government (Jacobs 1993: 4-5, Baker et al. 2019: 4-5). The British NHS Act of 1946 was "formally applied to England and Wales" and subsequently incorporated in Scotland and Northern Ireland (Jacobs 1993:4-5) in 1947 and is the oldest "state-based healthcare system in the world and refers to the publicly funded, comprehensive healthcare system in the United Kingdom (Baker et al. 2019: 4-5, Delamothe 2008). The NHS was created in part in response to address public views and support for "major health care reform", including support for "hospital and community-based care" and "state-involvement" (Jacobs 1992: 215-17). The founding principle of the NHS is that healthcare should be free for all and this continues to hold true, alongside current NHS aims of supporting health and care generally, improving health and care service outcomes and efficiency, and assisting frontline service (NHS 2021). The NHS is funded primarily through taxation and went through several major controversial reforms, most recently in 2012. The NHS is a major facet of life in the United Kingdom; a recent poll on what makes people proud to be British included the NHS as the second most popular institution in 2018, and the most popular institution in 2020 (Smith 2018, Newton 2020). It is markedly different from either WHO or the CDC as the NHS includes all health efforts and care under this coordinated umbrella through the British government (Smith 2018, Jacobs 1993, NHS 2021).

The relationship between culture, historical events, and lexical changes is an essential area of study in linguistics and corpus linguistics (Baker 2011, Stubbs 1995). Because of this, it is important to take into consideration the changing circumstances throughout the year 2020, as these certainly played a role in communication choices and the data in these corpora. On 30 January, the WHO declared that a novel coronavirus outbreak had become a "Public Health Emergency of International Concern". It was not until 11 February that the WHO officially recommended the name of this virus be COVID-19, an acronym for coronavirus disease 2019. By

the end of February, cases worldwide, including in Iran and in Italy, were growing exponentially and on 11 March the WHO officially declared COVID-19 a pandemic. Later in March the CDC officially advised the public that no large gatherings occur, and many countrywide lockdowns and travel bans also happened in March and in April. Other notable events include cases in Africa surpassing 200,000 in June, and global deaths exceeding 800,000 on 22 August and 1 million on 28 September. The UK reentered lockdown on 5 November and the US shortly after topped 10 million infections on 8 November. In December, COVID-19 vaccines were authorized for use, including the Pfizer vaccine (first in the UK) and the Moderna vaccine. Although the pandemic continues into 2022, the following analysis focuses only on corpus data in 2020.

Previous studies utilizing Twitter as data sources span subfields including corpus linguistics, computer science, computational linguistics, and others. This also overlaps with computational sociolinguistics, a growing subfield dedicated to understanding language variation and change over time. Related studies focus on genres and subtopics within social media, including a variety of important applications like public health understanding and messaging. Work highlighting these goals includes studies on tweets concerning the Zika virus (Gui et al. 2018), mental health, obesity, and depression (Jo et al. 2013, Smith et al. 2016, Broniatowski et al. 2013, Mowery et al. 2017). Twitter itself is a widely popular social media, with outreach and interaction being major goals of the platform itself (Dunder et al. 2016, Pano & Kashef 2020). Gui et al. study tweets on Zika from different public health sources including the CDC. They point out that "social media have become a powerful channel for information seeking and sharing, especially in times when timely information is critical" like public health crises (2018: 820). Their study focused specifically on tweet polarity and found that the most popular tweets related to Zika were in the humor and joke category, rather than policy, infection, or research updates (2018: 825). Gui et al. also found an overall "disconnection between the [wider] public's understanding of a public health situation and available, scientific information sources provided by public health agencies" (2018: 820) and argue that the spread of rumors and misinformation

only further emphasizes this gap (2018: 828). Culotta et al. 2010 focus on CDC tweets and reports in order to analyze the results of computational models applied to predicting rates of flu and flu-like illnesses in a population (115). King et al. 2013 investigate the "role of Twitter in informing, debating, and influencing opinion" in health policy with the passage of the Health and Social Care Bill in the United Kingdom. By applying sentiment analysis with a corpus of tweets related to health reforms, they found that Twitter often operates "as a meeting place in which those on one side of the argument come together to share information and to reinforce their own views" (2013). Other relevant literature involves the use of social media data for tracking public understandings and information regarding health and disease (Jo et al. 2013, Smith et al. 2016, Broniatowski et al. 2013) and is particularly applicable to the chapter on twitter COVID-19 messaging.

Additional studies focus on Twitter in terms of linguistic variation in different registers and dialects of English. Einsenstein et al. (2014) analyze a large dataset of tweets from 2009-2012 to examine lexical diffusion across the US. They find evidence of "demographic similarity" providing a prominent role in "linguistic influence" and argue further that this dataset evidences spoken American English dialectal patterns and variation (Eisenstein et al. 2014). Goel et al. examine linguistic variation on Twitter with a geotagged dataset from June 2013-2014 to detect how "changes propagate through social networks" (2016: 42). They find that linguistic changes on Twitter are a "form of information diffusion across social networks", with "no evidence to support the hypothesis that geographically local ties are more influential" than other types of connections within networks (Goel et al. 2014: 54). Grieve et al. (2018, 2019) use Twitter as a major source of data in this area for a variety of relevant work in understanding English language variation across different locations and communities, including the US and the UK. Clarke and Grieve analyze context and stylistic variation in Donald Trump's Twitter account between 2009 and 2018 utilizing multivariate analysis of grammatical patterns (2019). They find evidence of specific tweet styles including "conversational, campaigning, engaged, and advisory discourse"

and note the importance of his "intended audience" (2019). Grieve et al. (2018) apply Twitter as a corpus to find changes and introductions of new "words on American Twitter," with data extracted from 2013 and 2014 to map and understand the use and locations of these words (2018: 294-5). They argue that there are five distinctive regional areas and "patterns of lexical innovation" (293) and discuss the implications for the way these features are spread and shared through Twitter (2018: 300;,313). Major findings include their argument that "cultural geography, rather than physical distance or population density, is the main determinant of regional patterns of lexical innovation" (311). Similar methods are utilized on British Twitter data in comparison with the BBC Voices project to understand lexical innovation in the UK (Grieve et al. 2019: 17), and they find that "patterns of regional linguistic variation are relatively stable" in both datasets (2019: 16). These studies demonstrate the wide range of utility and opportunities for incorporating Twitter data as corpora.

In addition to the proliferation of studies on Twitter data, work is currently being done on COVID-19 and social media and specifically analyzing the COVID-19 twitter dataset (Banda et al. 2021; Lamsal 2020), which contains tweets on COVID-19 from throughout the pandemic. This dataset includes many different languages, but the most frequent tweets are in English, Spanish, and French (Banda et al. 2021). Many of these studies utilize machine learning and computational methods from natural language processing (Wicke and Bolognesi 2020), while Semino and Koller (2020) focus on much smaller datasets and use qualitative methods and analysis of keywords in context to understand differences in metaphors describing COVID-19 and related terms such as *lockdown* and *pandemic*. Additional research on COVID-19 related tweets (Semino & Koller 2020, Semino 2021, Wicke and Bolognesi 2020) focus on smaller datasets with a variety of different methods. Wicke and Bolognesi use topic modeling to investigate a corpus of tweets containing the phase *#COVID-19* from March- April 2020; they find evidence (corroborated by Semino and Koller 2020) of different metaphor groups in use in describing COVID-19 and the pandemic's effects, including metaphors relating to "war-framing,

monster, storm, and tsunami". They argue that the war-framing metaphor is predominantly used in their dataset (2020).

3. Methodology

The goal of this chapter is to examine and uncover differences and similarities in messaging choices by the CDC, WHO, and NHS, through corpus-based methods. Specifically, trigrams are analyzed over time to discover linguistic patterns throughout 2020, for each source and context of usage. In particular, what multiword units were increasing, decreasing, or remaining stable in use over time, relative to others in each corpus is of interest. Stability of trigrams and influence of external events are also taken into consideration. In order to accomplish this analysis, each corpus is divided into three-month subcorpora for initial analysis and inspection of results: January- March, April-June, July-September, and October-December of 2020. Keyword analysis is also studied comparing each governmental unit against the other two groups across 2020. Keywords are based on "frequency difference" (Gabrielatos & Marchi 2012) and index "aspects of content characterizing" each group. Keyword analysis is conducted using R packages tidytext and quanteda (Silge & Robinson 2017, Benoit et al. 2018) with contingency tables; loglikelihood values (G2) are calculated for each keyword. Keywords are useful for revealing "trends across time" (Kretzschmar et al. 2004: 40-41) and indicate differences in communication by the CDC, WHO, and NHS. Keywords also often underscore grammatical and lexical characteristics of the texts being analyzed (Culpepper 2009: 43-44). To address and determine the contextual usage (Hoey as quoted in Baker 2010: 329) of top keywords, inspection of concordance lines and keywords-in-context (kwic) is conducted and discussed for the CDC, WHO, and NHS.

Before testing and running scripts to obtain the Twitter data, official organization publications and websites were checked to ensure that each account included was officially

associated with that organization[29]. Although these organizations used other languages than English in their communication, English-only tweets were included for analysis. A full breakdown of each corpus, subcorpus, and account is presented in the Appendix. Next, python (verson 3.7.4) was used with libraries pandas (version 1.2.3, McKinney 2010) and twint (Zacharias 2020) to scrape all tweets from each account from 1 January 2020-31 December 2020[30]. This script went through multiple versions and testing to ensure that the selected accounts were all actively tweeting during this time period, and that they did discuss COVID-19 and/or the coronavirus. This script was then updated and re-run at various stages of the research process to continue obtaining corpus data. After obtaining the full dataset for all tweets from 2020, the data was reformatted and encoded for use with CQP and polmineR using an R script and the command line (Evert & Hardy 2011, Blätte & Leonhardt 2019); this script included tokenization, appending corresponding metadata for each token, keeping special characters like the # and username symbol @, and checking the tweet ids to ensure that no tweets were repeated. The tweets from each account with corresponding metadata including date of tweet, organization, month, and year, were stored in separated csv (comma separated values) files in corresponding folders for each governmental organization. Additional linguistic attributes were appended, including part of speech tags, lemmas, and sentence level annotations, added in the final processing step for CWB-encode. For a more thorough discussion of the methods and of the corpus compilation process, see chapter 3.

4. Analysis & Discussion: Trigrams

An important variant of frequency analysis is the investigation of multiword units, sometimes called lexical bundles, ngrams, sequences, or formulas in a corpus; this can be broken down into

[29] For each organization, there were multiple parody accounts or other types of Twitter accounts that had names similar to the CDC, NHS, or WHO but were not actual official accounts. These were purposefully excluded from this study; however, they would be worth considering for future work.
[30] The full corpus also includes tweets from 2019; however, when inspecting the data, little and sometimes no mention of COVID-19 occurred.

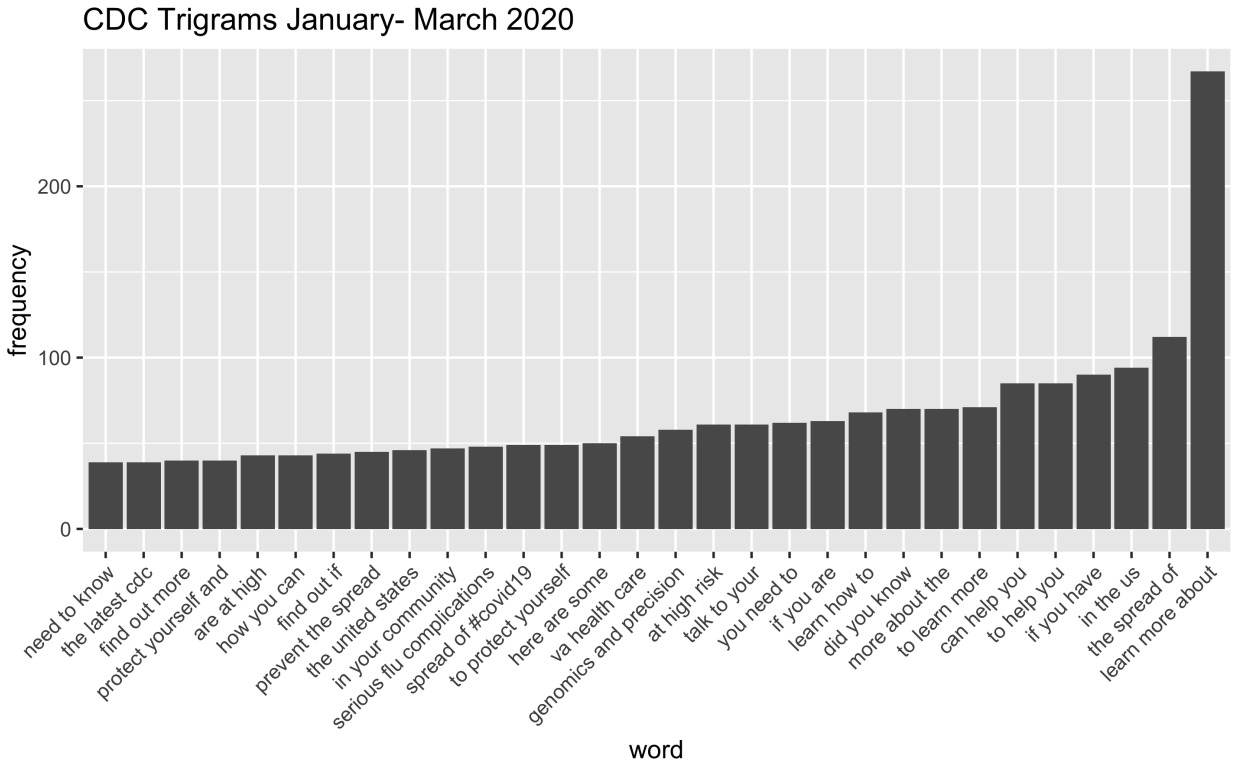
unigrams (one token), bigrams, trigrams, and so on. The top trigrams for each corpus, shown below, reveal what each governmental organization chose to emphasize through their linguistic and rhetorical choices. Investigating patterns of stability and of variation over time also highlights the changes in focus and changes relative to events and updates during the pandemic.

Trigrams were selected for this analysis because of their importance in texts. Previous work has shown that trigrams hold various important functions in communication, including expressions of stance or attitude, referential functions, discourse organizing functions, and topic introductions or expansions (Biber et al. 1999, Baker 2011, Biber & Barbieri 2007). Trigrams were also selected because of their role in exposing and revealing larger discourses within language (Hunston 2010, Römer 2012). Stubbs explains this well, arguing that "culture is encoded not just in words" but in sequences of words and different word combinations (Stubbs 1995).

In order to accomplish analysis, each corpus is divided into four time groups: January- March, April- June, July- September, and October- December. The subcorpora sizes differ[31], and all trigrams went through the same compilation and organization procedures. First, all trigrams were obtained for the specified time samples. Following this, all tokens were collapsed by case into lowercase tokens, and extra punctuation was removed. Next, a stopword list algorithm was applied to remove only trigrams in which all three tokens were stopwords, like *the and a.* The results shown below include only the top thirty trigrams in each time sample.

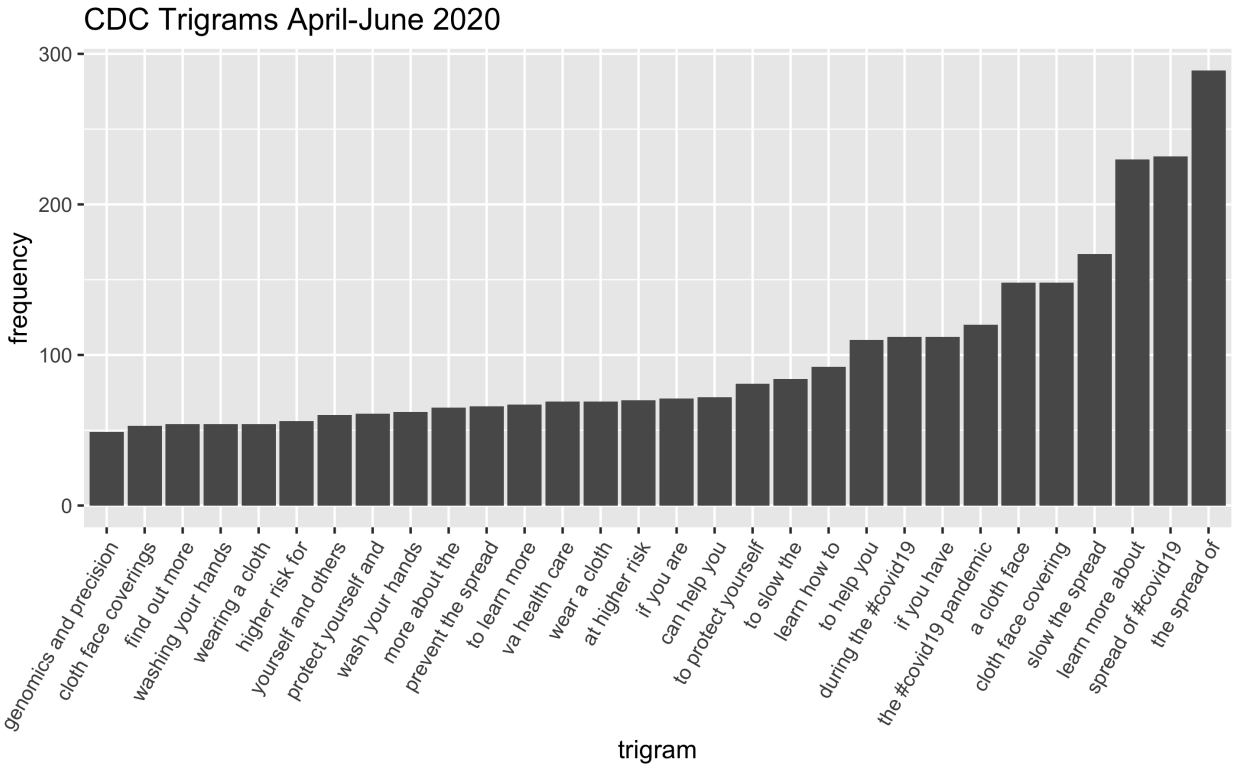
[31] A full breakdown of these sizes is available in Appendices H-J, pages 170-172.

Figure 6.1



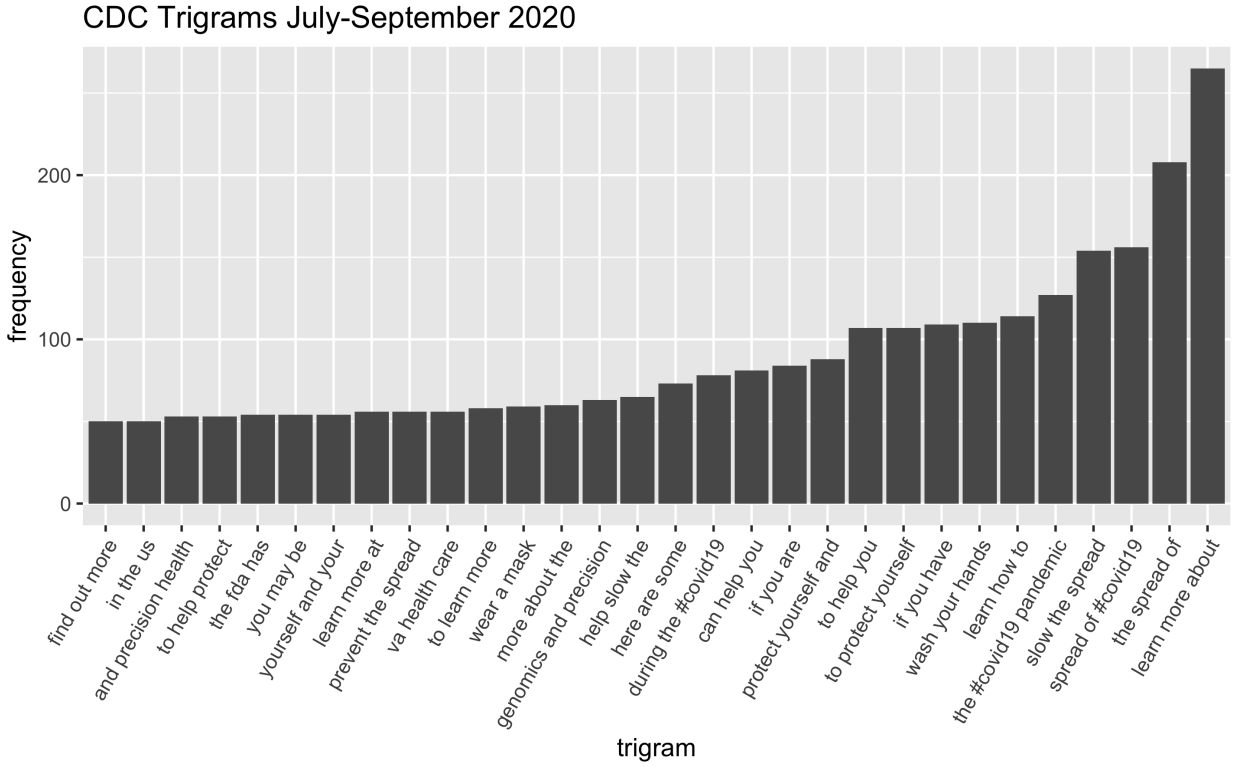
At the beginning of 2020, despite the fact that COVID-19 was spreading rapidly throughout the world, mentions of COVID-19 were few in the American media, especially in January and not until March. This is visible and reflected in the frequency distribution of trigrams shown above in Figure 6.1. The only mention of COVID is present in the trigram *spread of #covid19.* Other top trigrams at this time reflect a US-internal focus and a focus on individuals: *in the US, if you have, to help you, can help you,* and *talk to your.* Although some of these multiword units remain highly frequent throughout 2020, others decrease in use, as shown in Figure 6.2 below.

Figure 6.2



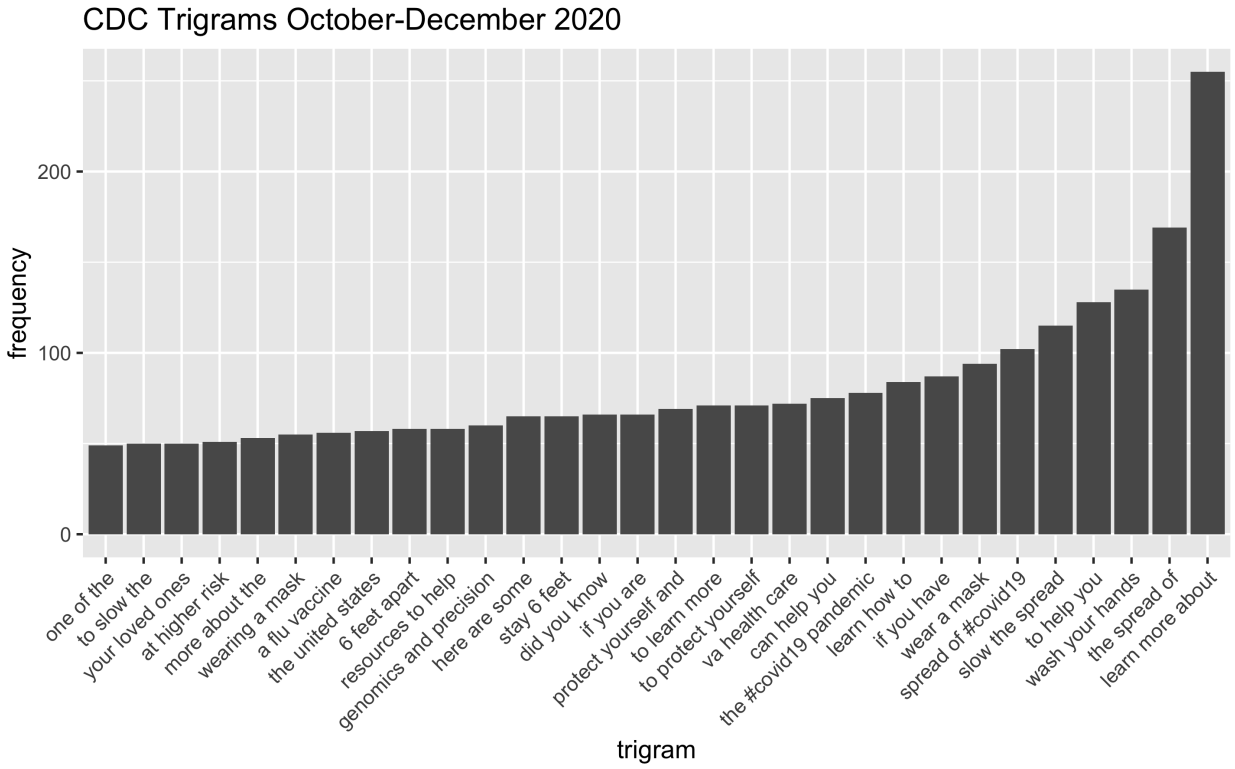
This time group is especially interesting considering the circumstances: especially in April, cases were rising exponentially in the US, and much work was being done across the world to increase understanding of the virus and to encourage individuals to be careful (Schuchat & CDC COVID-19 Response Team 2021). *Spread of #covid19* rises in frequency, and new tokens are introduced including *slow the spread, cloth face covering, wear a cloth, washing your hands, during the #covid19* and *the #covid19 pandemic.* Trigrams including *the spread of, learn more about, if you have,* and *to help you* remain stable in use. Figures 6.3 and 6.4 show additional examples of stability and change present in the CDC accounts.

Figure 6.3



Beginning in July, top trigrams are similar to the previous time group, such as *learn more about, the spread of, spread of #covid19,* and *slow the spread.* Others rise in frequency including *the #covid19 pandemic, learn how to,* and *wash your hands.* Additional advisory phrases are *prevent the spread* and *wear a mask,* and references to individuals and the US are also present in the frequency distribution above. *In the us, yourself and your, if you are, to help you* are a few examples.

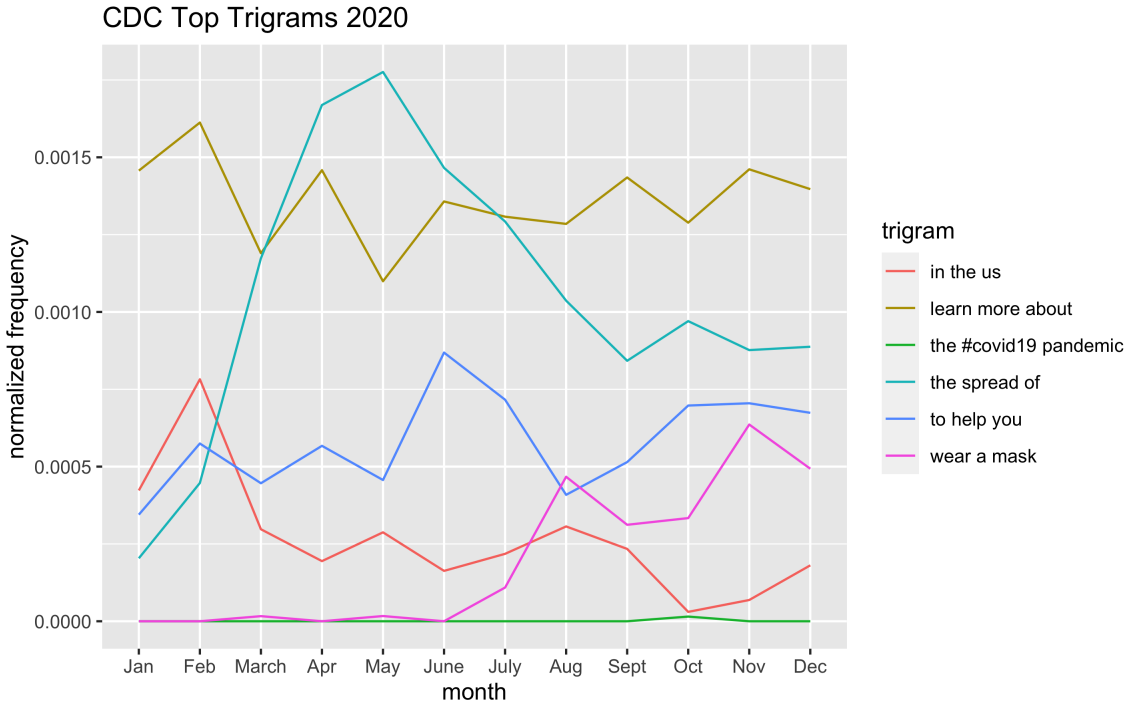
Figure 6.4



Throughout 2020, several key trigrams remain highly frequent, including *spread of #covid19, the spread of,* and *learn more about.* Explicit mentions of cloth face coverings also remain prevalent until October 2020, when trigrams like *wear a mask* and *wash your hands* rise in frequency. Other trigrams in this time sample continue references to individuals like *your loved ones* and *to help you.* References to information, resources, and recommended actions are also present in the frequency distribution above.

Figure 6.5 illustrates the variation present in the CDC accounts throughout the year. Each of the frequency distributions reflect characteristics of the CDC accounts including a rise in specific recommended actions, a focus on Americans as individuals and on US-internal concerns, and references to information about the pandemic or related health topics.

Figure 6.5



The CDC trigram results provide some interesting insights into the CDC's communication via Twitter throughout 2020. The CDC accounts increase in use of advisory phrases, such as *wash your hands* and *wear a mask*. Beginning in April, trigrams *a cloth face* and *cloth face covering,* and *wear a cloth* are introduced as the most frequent items. These remain stable and of high frequency in July- September but completely fall out of the top frequent trigrams later in the year. It is possible that these terms were selected instead of explicitly using the term *mask* because of concerns about the lack of medical grade N95 and KN95 masks for healthcare workers; it could also be connected with the highly politicized nature of mask wearing in the US. The chief medical advisor to the Biden Presidential Administration and director of the U.S. National Institute of Allergy and Infectious Disease Dr. Anthony Fauci has repeatedly pointed out that everyone should wear masks and that "it should not be a political issue. It is purely a public health issue" (Aratani 2020, Douthat 2020). All cloth covering references in the

CDC peak in use until October-December of 2020, when most references to masks include *wear a mask* and *wearing a mask,* rather than referring to masks as cloth coverings.

The semantic categories below in the summary for each account are not exclusive but are grouped according to contexts of use. Some trigrams also overlap in categories, like *you need to* and *in your community* fit into the group individuals and people but are also general information.

Table 6.1: CDC Summary of Trigrams

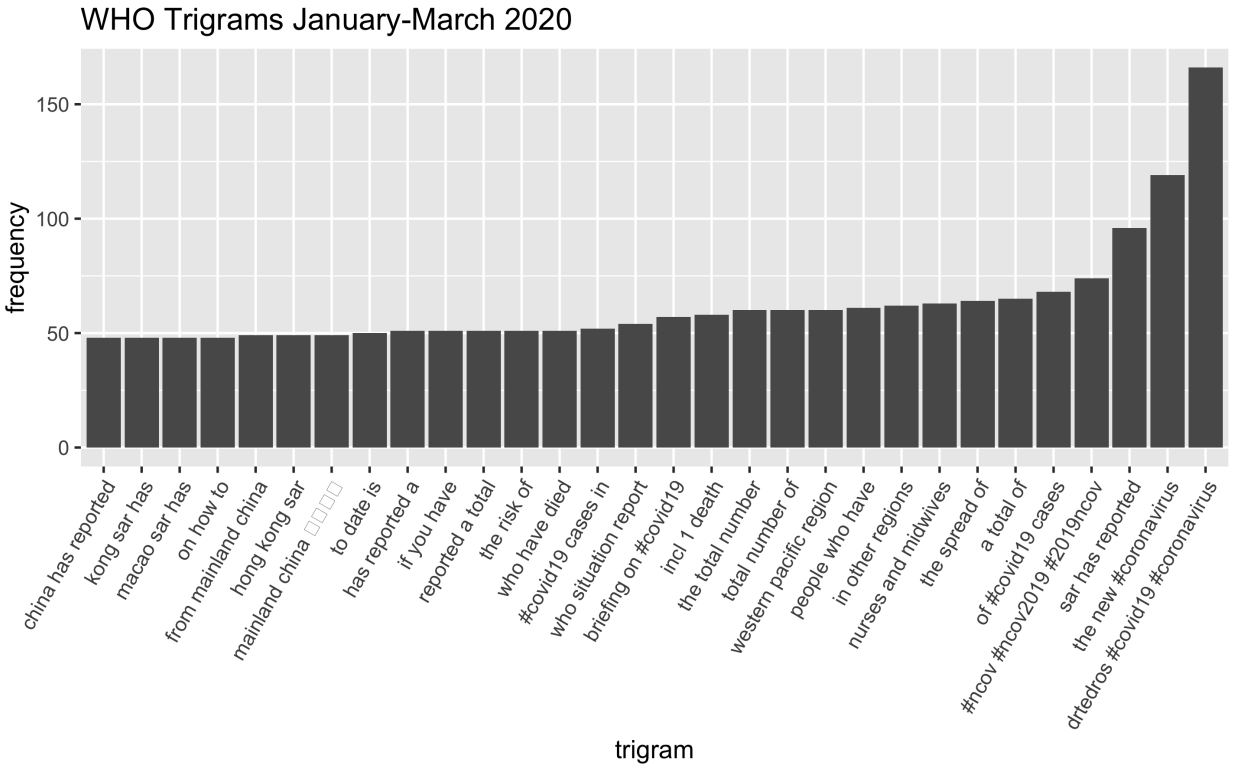
Semantic Category	**Increased**	**Decreased**	**Stable**
Individuals and People		if you have you need to how you can to protect yourself talk to your in your community	did you know to help you protect yourself and
Regions, Countries, and Organizations	the United States	genomics and precision va health care in the US the latest cdc	
Cases and Deaths			
General Information	at higher risk a flu vaccine** spread of #covid19 the #covid19 pandemic	more about the a total of need to know	learn more about learn how to find out more/prevent the spread(through September) to learn more
Advisory phrases	wearing a mask washing your hands slow the spread wash your hands wear a mask stay 6 feet wear a cloth	cloth face coverings wearing a cloth	the spread of

The CDC trigram summary emphasizes the contexts of use and themes present over the year. Stable trigrams like *did you know* and *to learn more* emphasize the focus on individuals as an audience. Advisory trigrams increased in use like *the United States, washing your hands,* and *stay 6 feet,* in addition to consistent references to general information related to the pandemic.

Stable general information trigrams throughout 2020 include *learn how to, find out more, learn more about,* and *to learn more.*

The trigram results for the WHO accounts follow the same criteria and divisions over time. Figure 6.6 displays the results for the frequency distribution from January- March of 2020.

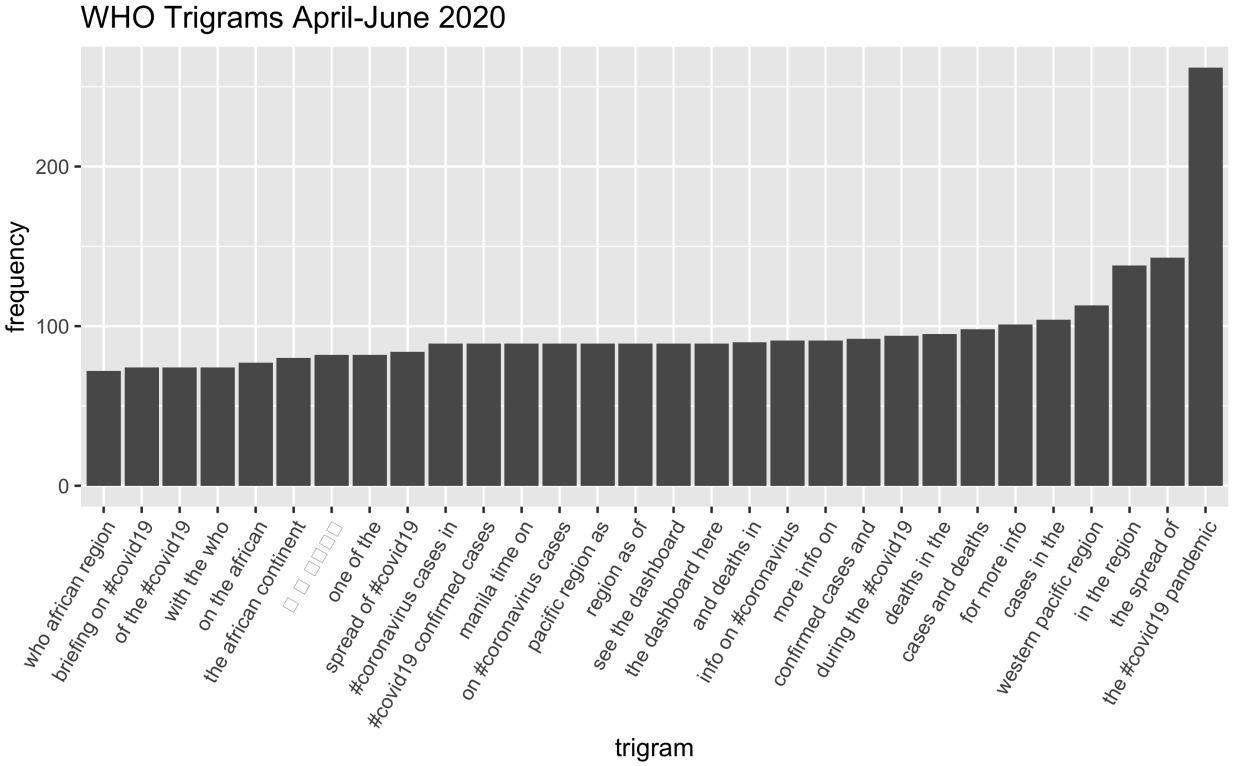
Figure 6.6



A variety of names and references to COVID-19 are used by WHO accounts exemplified in these trigrams: *the new #coronavirus, #coronavirus #ncov2019 #2019ncov,* and *of #covid19 cases.* In addition to this, the WHO tweets tend to mention different regions and locations in numerous trigrams: *china has reported, sar*[32] *has reported, from mainland china, western pacific region.* References to numbers of cases and information are present in trigrams like *the total number, who situation report,* and *to date is.* Very few trigrams refer to individuals or people, aside from the following examples: *nurses and midwives, people who have,* and *if you have.*

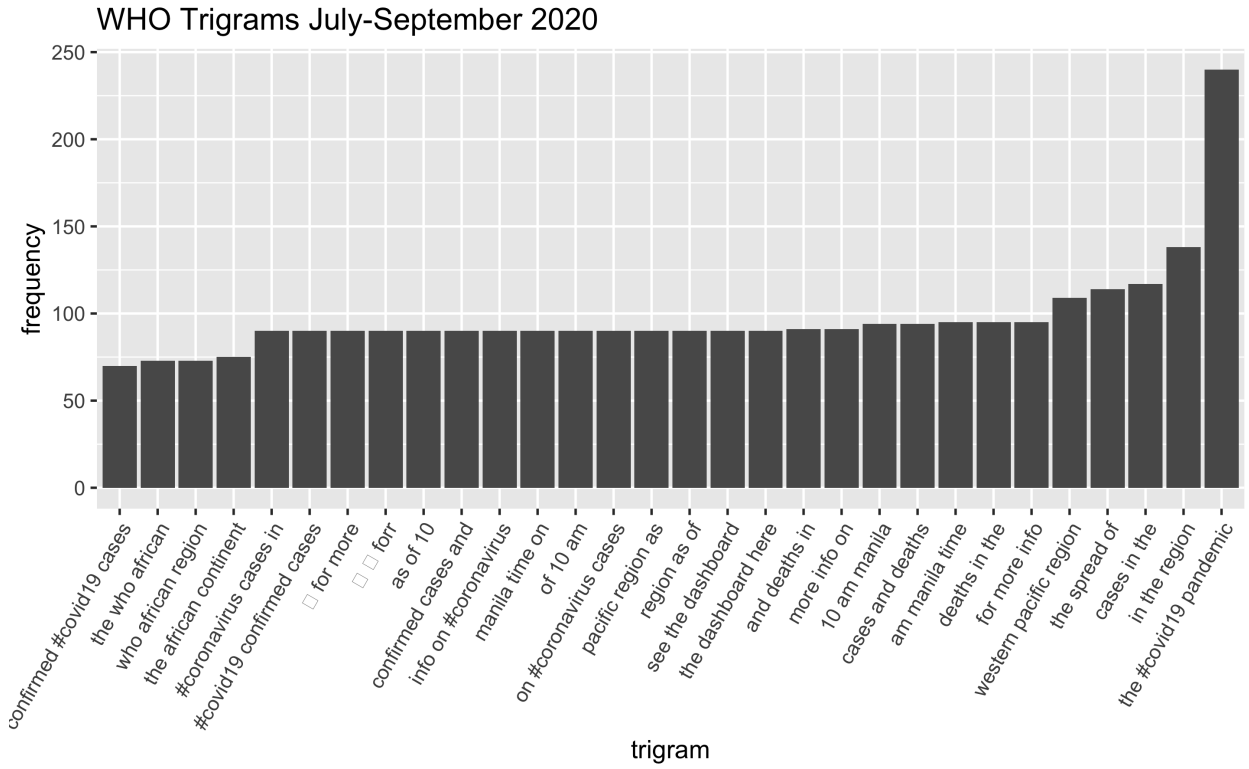
[32] SAR refers to Hong Kong Sar, which is a special administrative region of the People's Republic of China (HKSAR 2014).

Figure 6.7



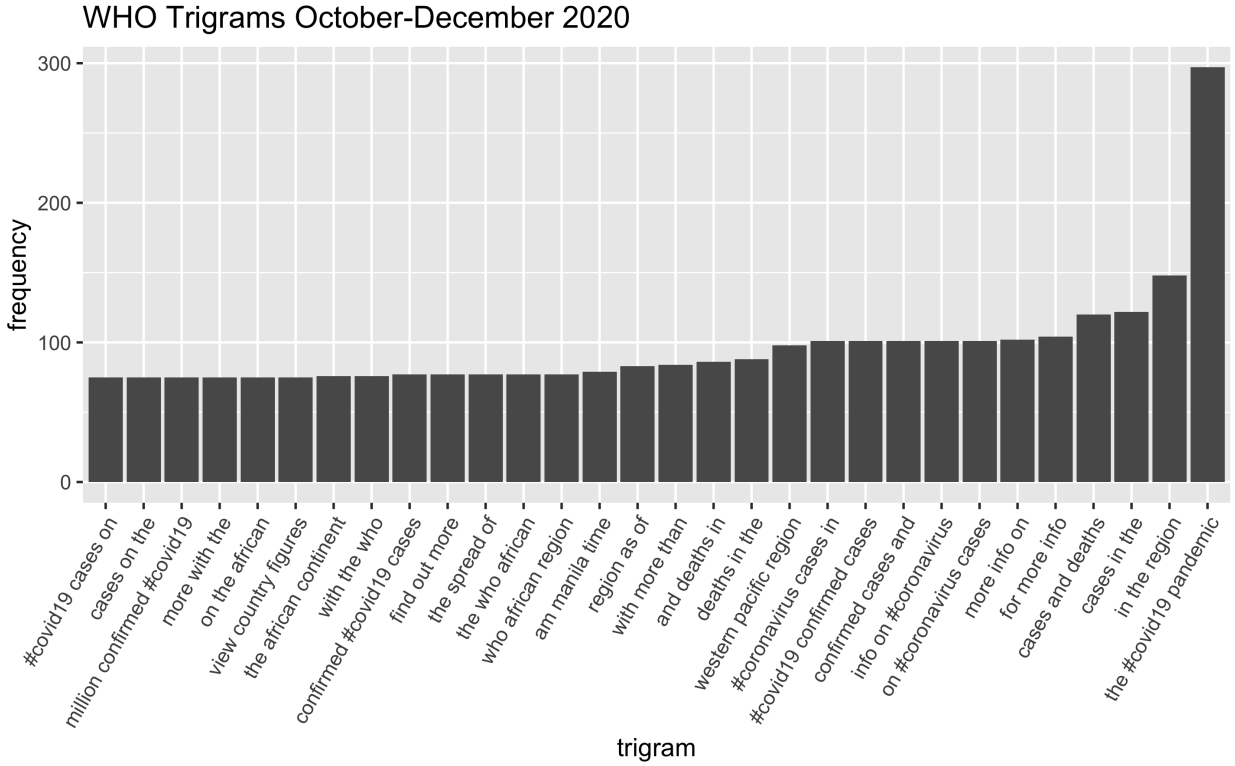
In April-June, broader themes and discourses persist, such as the WHO's tendency to emphasize regions and to reference COVID-19. Trigrams like *in the region, cases and deaths,* and *pacific region as* are just a few examples of these themes. In addition, references to cases and numbers rise, shown in trigrams like *cases in the, cases and deaths, confirmed cases and.* The trigram *western pacific region* remains highly frequent, along with *the spread of* and *during the #covid19. The #covid19 pandemic* rises in the frequency distribution to be the top trigram by WHO accounts at this point in time.

Figure 6.8



In July- September, shown above in Figure 6.8, *the #covid19 pandemic* remains highly frequent, as do trigrams *in the region, cases in the, the spread of,* and *western pacific region.* References to cases and regions are still used, including different trigrams like *who African region* and *the African continent.* The latter trigrams might have risen in frequency at this time due to the rise in COVID-19 cases, in Africa surpassing 200,000 earlier in June. The trigrams *info on #coronavirus* and *more info on* remain stable into this time frame, but *#coronavirus cases in* and #covid19 confirmed cases decline in frequency.

Figure 6.9



In the last time group, the WHO accounts include more varied references to COVID-19 *the #covid19 pandemic, on #coronavirus cases, info on #coronavirus, #coronavirus cases in, confirmed #covid19 cases,* and *million confirmed #covid19.* This is particularly interesting considering WHO recommended the name of covid be COVID-19 earlier on in the year. The trigrams include little to no references to individuals, people, or advice, and instead discuss case and death numbers, regions, and COVID-19 itself. Stability in some trigram units occurs, like in *the #covid19 pandemic, in the region, more info on,* and *cases in the,* which remain highly frequent through the year's end.

Figure 6.10

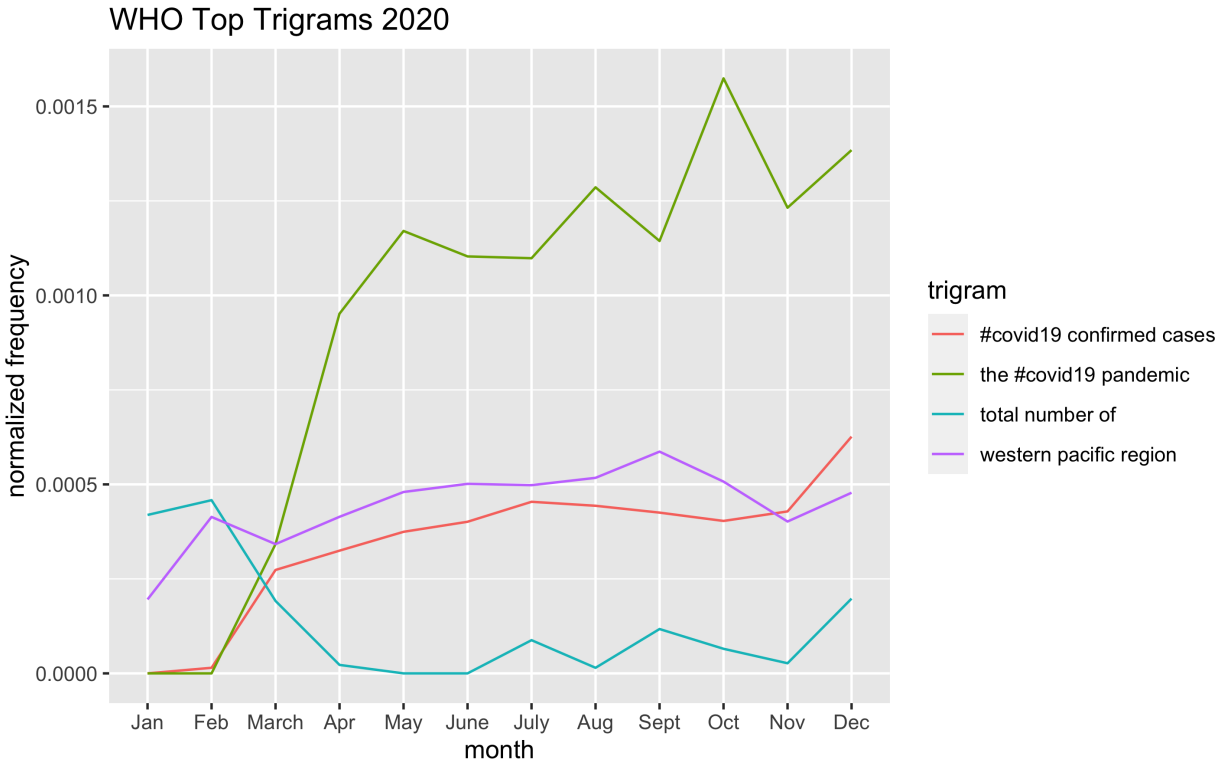


Figure 6.10 displays the top trigrams present throughout the year and their variation in use, relative to the frequencies of tokens per month. References to *#covid19 confirmed cases* peaks in December and use of trigram *the #covid19 pandemic* peaks in frequency in October. *Total number of* is used most often in February and *western pacific region* in September. Table 6.2 provides an overall summary of the trigrams in the WHO accounts.

Table 6.2: WHO Summary of Trigrams

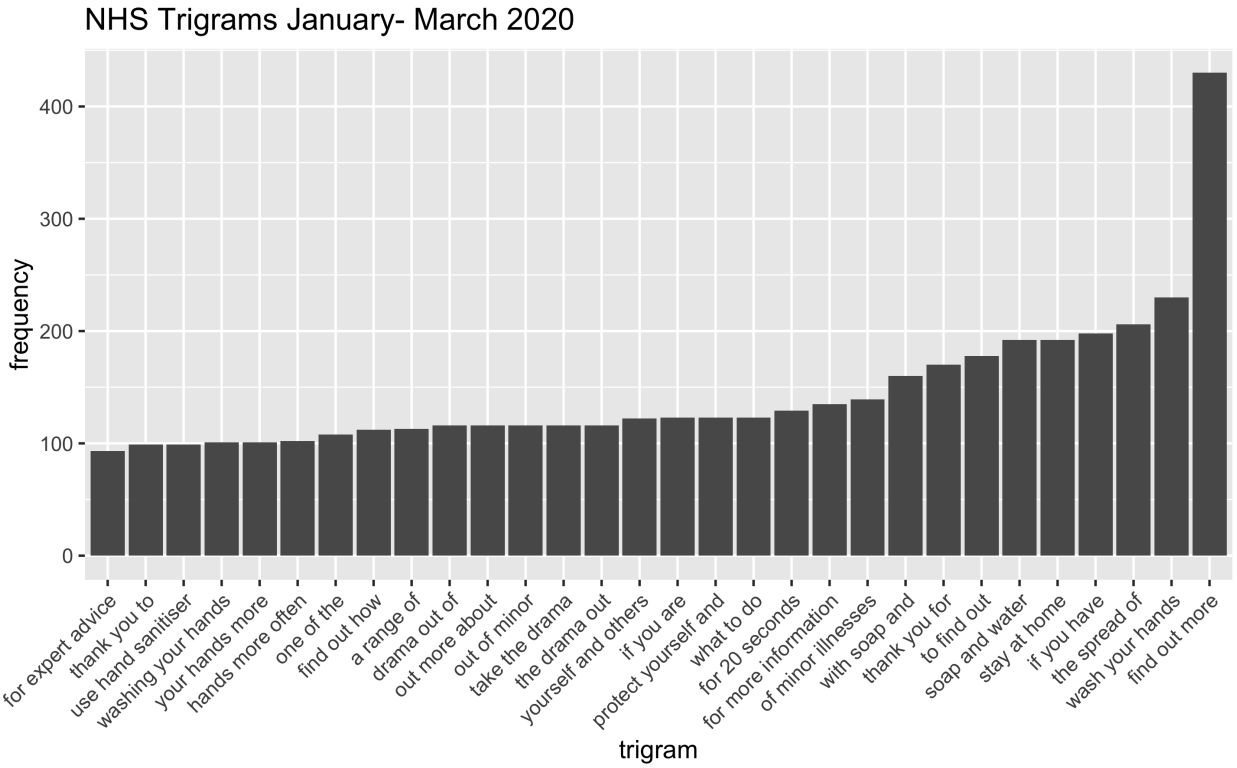
Semantic Group	**Increased**	**Decreased**	**Stable**
Individuals and People		if you have people who have nurses and midwives	
Regions, Countries, and Organizations	in the region region as of	WHO advisory report in other regions China has reported sar has reported from mainland China	western pacific region in the region

Cases and deaths	view country figures deaths in the #covid19 confirmed cases #covid19 cases in cases and deaths deaths in the cases in the confirmed cases and	a total of the total number	
General Information	for more info more info on	on how to to date is #ncov #ncov2019 #2019ncov the new #coronavirus	the spread of for more info the #covid19 pandemic briefing on #covid19
Advisory phrases			

The summary of WHO results highlight evidence of a predominant focus on case information, COVID-19, and different regions. They include links to the WHO COVID-19 dashboard, a resource on the WHO website dedicated to information about COVID-19 case numbers across the world (WHO 2021). Trigrams like *more info on, in the region, western pacific region,* and *the #covid19 pandemic* emphasize these broad trends. The decrease in use of trigrams like *if you have, people who have,* and *nurses and midwives* highlights a key distinction from the CDC and NHS and a lack of attention to individuals or advice for individuals. The majority of the WHO's trigrams fit into general information and references to cases, deaths, regions, and countries, emphasizing the WHO's overall focus on global reporting.

The NHS official accounts are analyzed below. Their tweets reflect similar themes and interests as the CDC and the WHO accounts, including a focus on broader updates like the WHO and on individuals and on recommended actions like the CDC. The specific frequency patterns highlight key differences from the CDC and the WHO tweets.

Figure 6.11



The first subcorpus, from January to March includes notable top trigrams: *find out more, wash your hands, the spread of, if you have,* and *stay at home.* Unlike either WHO or the CDC, COVID-19 is not mentioned at the beginning of 2020 in the top frequency distribution. Instead, advisory phrases proliferate: *wash your hands, stay at home, protect yourself and,* and *use hand sanitiser. Soap and water* and *washing your hands* also fit in with these themes. An emphasis on individuals is also demonstrated in trigrams *if you have, if you are, protect yourself and,* and *thank you for.*

Figure 6.12

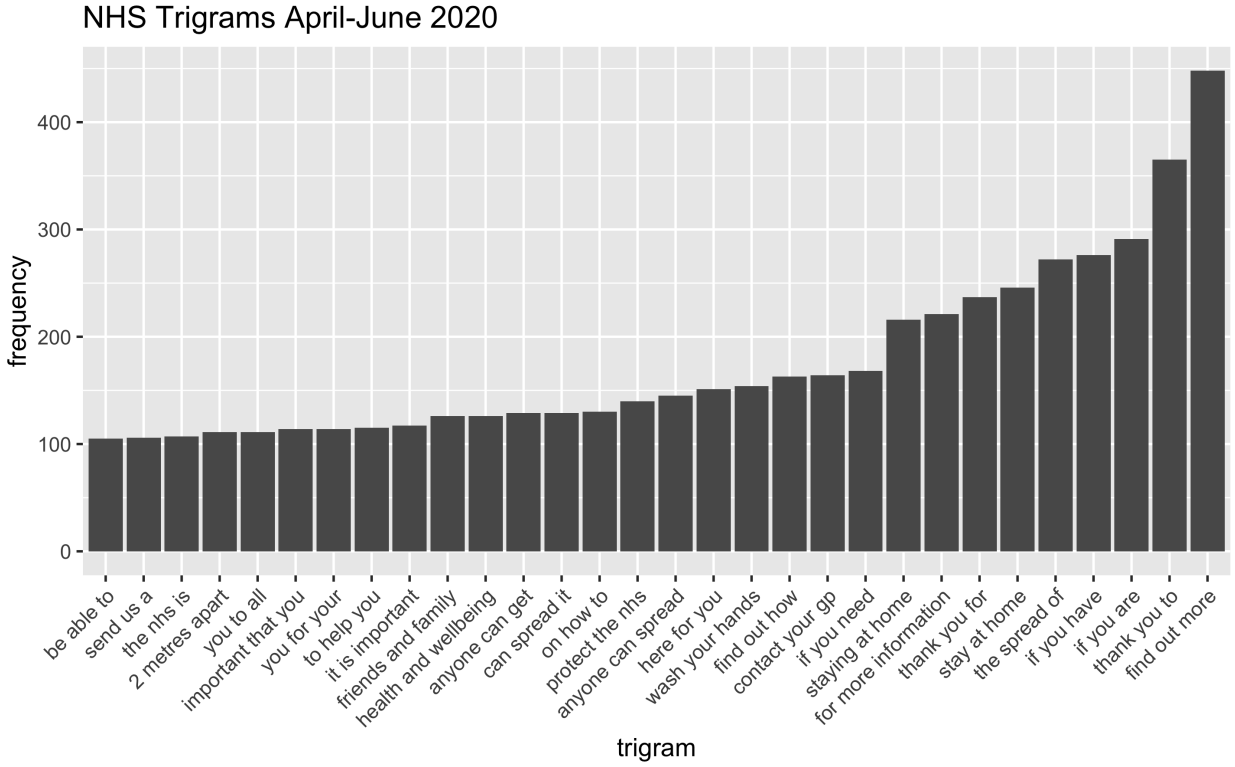
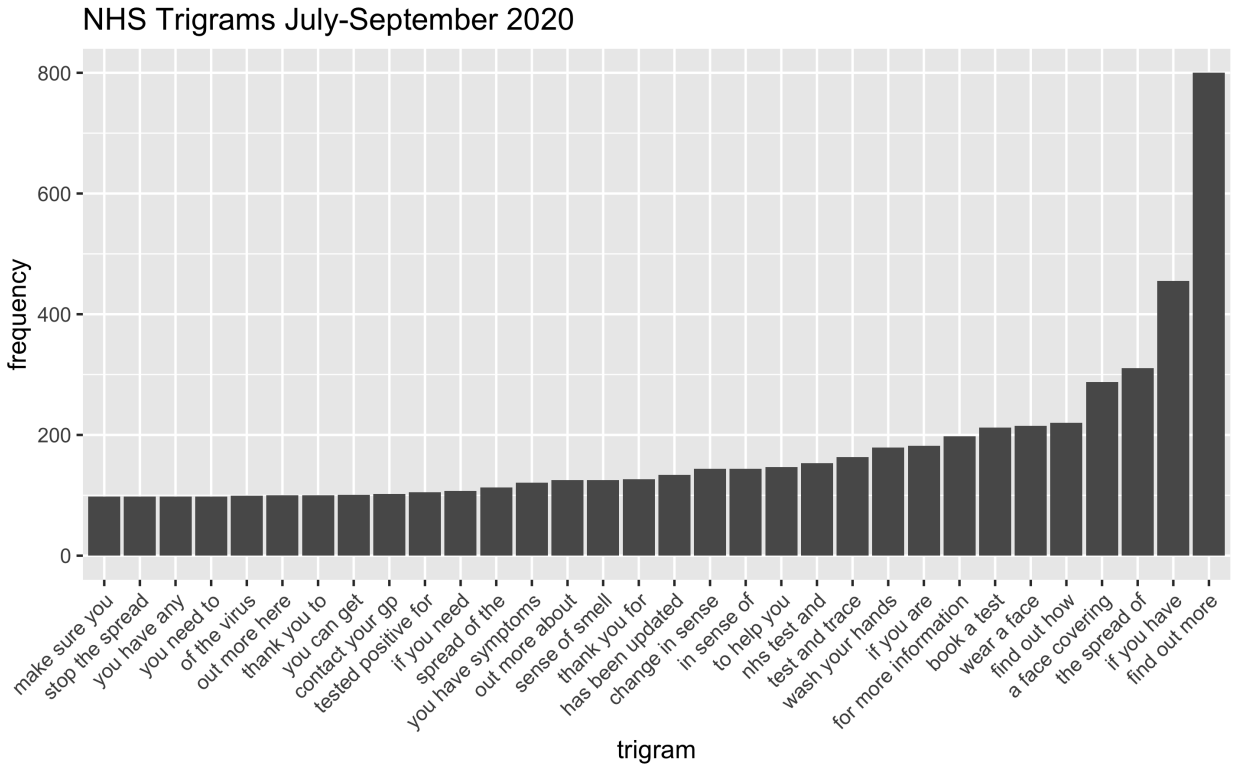


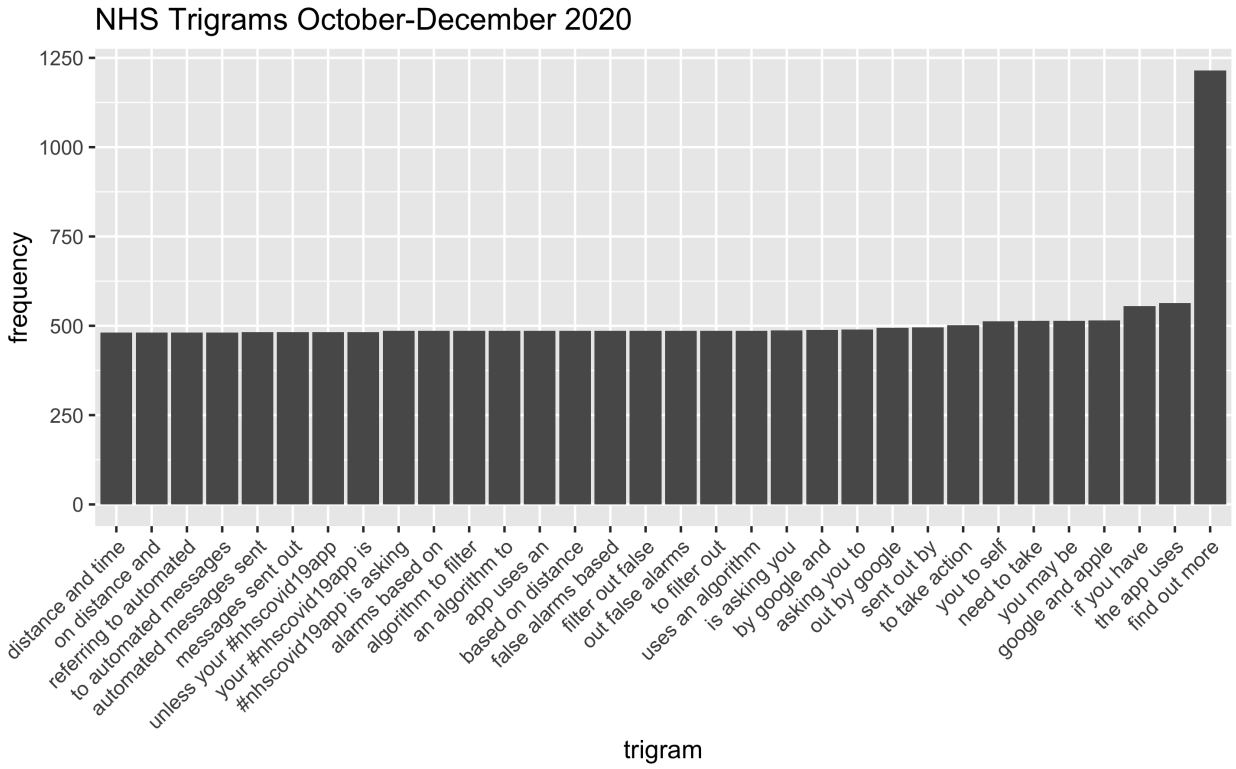
Figure 6.12 shows that stable top trigrams remain in frequent use by the NHS accounts: *find out more* and *the spread of.* More references to people, communities, and individuals occur in this time: *anyone can get, anyone can spread, friends and family,* as do references to lockdown (though without explicit use of the latter term): *stay at home* and *staying at home.* Other changes include a rise in "thank you" related trigrams: *thank you to* and *thank you for*, references to the NHS: *the nhs is* and *protect the nhs*, and other recommended actions like *contact your gp, stay at home,* and *2 metres apart.* References to COVID are not present in the top frequency distribution shown above.

Figure 6.13



The trigrams from July-September emphasize additional recommended actions and references to the NHS Test and Trace program: *wear a face, a face covering, stop the spread,* and *test and trace.* The NHS Test and Trace program was developed rapidly in response to the COVID-19 pandemic, and officially implemented in May 2020 (NHS 2021). The Test and Trace program was executed to alleviate issues in the pandemic including spread of infection: "we have introduced this service to help return life more to normal, in a way that is safe and protects our NHS and social care" (2021). The Test and Trace program includes individualized advice so that people who have been exposed to the virus or have symptoms access customized recommendations and support (NHS 2021). References to symptoms are also present and introduced at this time: *sense of smell, you have symptoms, change in sense.* The phrase *of the virus* is the first specific reference to COVID-19 in top trigrams utilized by the NHS. *For more information* is a stable trigram from the previous time sample.

Figure 6.14



Find out more remains the top trigram in use by the NHS throughout 2020. Many changes in the trigram frequency distribution are presented above. The NHS was focused on making sure the public knew about the recently and rapidly developed app, as part of the Test and Trace program (NHS 2021) and was communicating frequently about it. The app was officially developed and deployed to adults age 16 and over, free for download in England and Wales, on 24 September, and includes multiple tools and resources like contact tracing and alerts (NHS 2021)[33]. All advisory phrases are essentially replaced by references to the app at this time: *uses an algorithm, the app uses, your #nhscovid19app is, automated messages sent.* No other references to COVID-19 occur beyond the name included in the app.

[33] App development began in August as part of the NHS Test and Trace Program and is owned by the Department of Health and Social Care, through the efforts of a team of experts from NHS Digital, the Turing Institute, Oxford University, Accenture, NHSx, VMware Pivotal Lab, and Zuhlke Engineering, in coordination with the UK National Cyber Security Centre (NHS 2021).

Figure 6.15

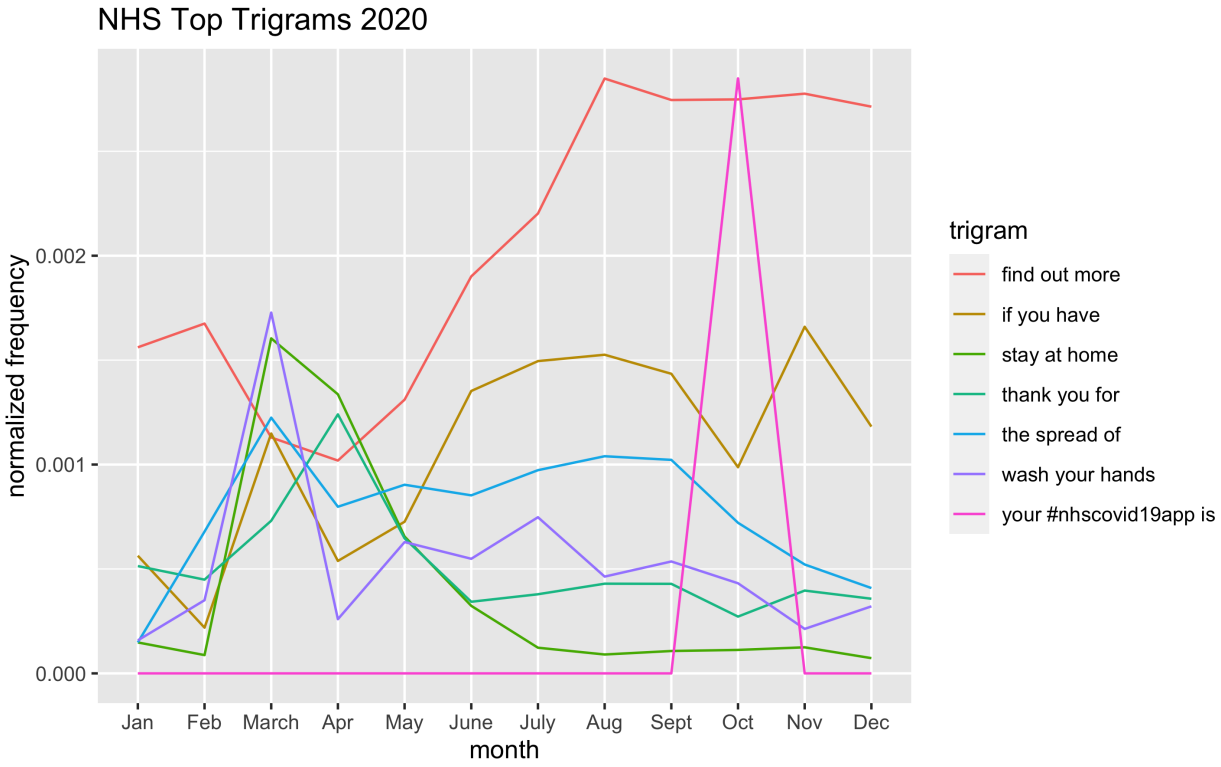


Figure 6.15 confirms the variation in top trigrams present across 2020. *The spread of* is used most often in March, as are *stay at home* and *wash your hands.* In November, *if you have* is used most often and references to the NHS app are very frequent in October, like the trigram *your #nhscovid19app is. Find out more* is consistently used most often. Table 3 provides an overall summary of the NHS accounts trigrams and illustrates that very few trigrams fit into the stable category.

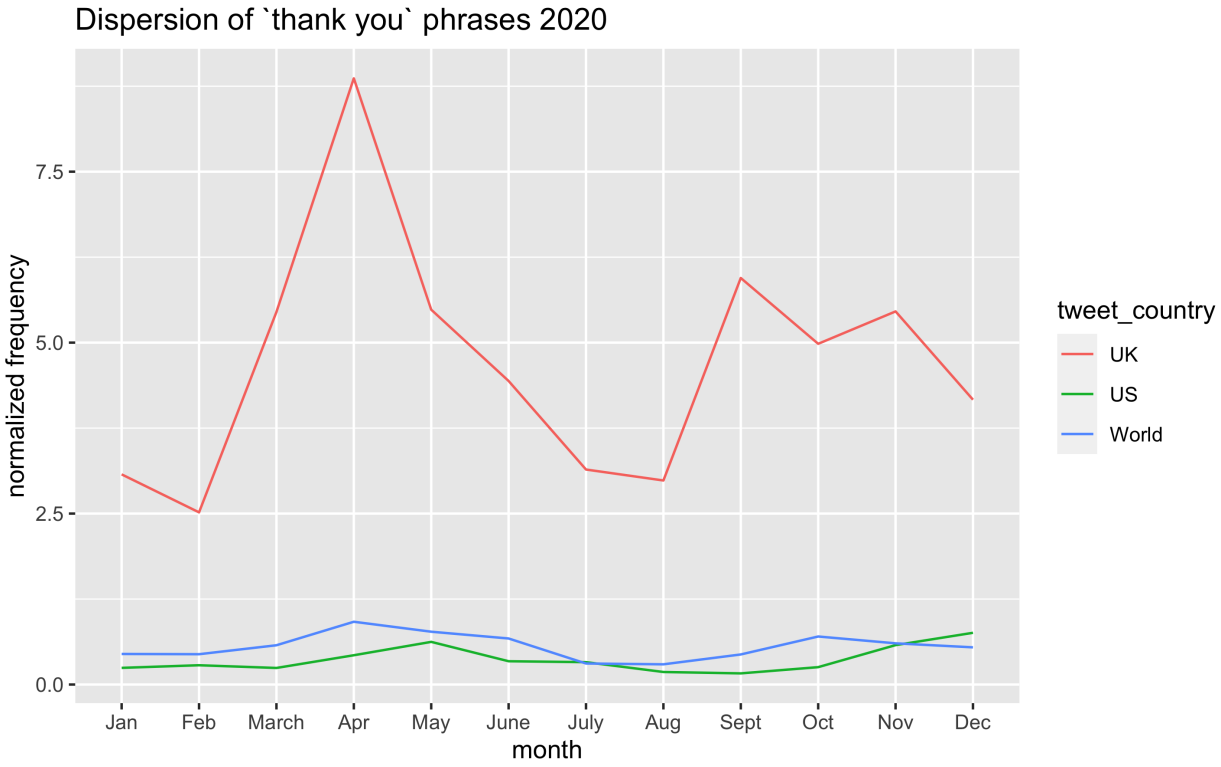
Table 6.3: NHS Summary of Trigrams

Semantic Group	**Increased**	**Decreased**	**Stable**
Individuals and People	if you need if you are to help you here for you you may be friends and family	thank you for thank you to your hands more	if you have
Regions, Countries, and Organizations	The NHS is		

	#nhscovid19app is asking (October) Google and Apple (October)		
General Information	anyone can spread anyone can get false alarms based	soap and water	for more information the spread of find out more find out how
Advisory Phrases	staying at home Wear a face Book a test Wash your hands It is important Stop the spread contact your gp*(July) a face covering*(July)	washing your hands use hand sanitizer stay at home take the drama out send us a protect the NHS For 20 seconds	Stay at home

The summary table above foregrounds the variation present in the NHS communication. The tweets include comprehensive actions, information, and resources, emphasized further by the development of the NHS Test and Trace Program and culminated in the NHS COVID app. The few instances of stable trigram units in the NHS accounts further underlines this distinction in their communication efforts on Twitter; however, these tweets include a variety of trigrams in many semantic categories, referencing individuals and people, general information, and advisory phrases. These all play a role in the diversity of topics and information covered and in providing comprehensive output to users, followers, and the public more widely. The NHS also includes many trigrams which refer to individuals and groups of people: *anyone can spread, anyone can get, friends and family,* and *to help you,* underlining an interest in promoting the health of individuals as well as communities. Each of these categories reflects the extensive goals and work of the NHS.

Figure 6.16



Overall, there are key differences in messaging between the CDC, WHO and NHS accounts.The CDC utilizes trigrams referencing individuals and people, advice, and information, and refers to Covid as COVID-19 or #COVID-19 throughout 2020. The CDC also refers to few key recommended actions, especially in the beginning of 2020 and in comparison to other tokens and trigrams. The CDC's messaging also changes, beginning the pandemic with fewer advisory phrases, and then adding additional recommended actions as the year progressed. Some examples include trigrams referencing face coverings which were replaced with references to masks in October-December.

The WHO uses more tokens that refer to regions, countries, and different cases and names of COVID-19. The WHO accounts focus exclusively on those topics, rather than including additional information or recommendations for health actions and protection for individuals, emphasized by the lack of trigrams in these categories. The WHO accounts also use emojis consistently and phrases with hashtags more often than CDC or the NHS. Overall the WHO is

focused on describing the broader situation of the pandemic across the world by describing and updating through case numbers and region-specific information.

The NHS accounts are much more varied, include wider ways of referring to people and key recommended actions, with most changes in communication catalyzed by the development of the app. The NHS communication is diverse and includes an assortment of recommended actions for individuals and groups of people, information about the virus and health, and access to additional information and links throughout the year. In this way, the NHS accounts were the most comprehensive out of each organization, by covering a variety of recommendations, information, and accessible additional resources. Very few frequent trigrams in the NHS actually reference COVID-19 itself, aside from the *virus* trigram in July-September and references to the NHS app. In October of 2020, the NHS twitter accounts focus mainly on communication regarding the NHS COVID-19 app; the trigrams from October-December of 2020 reflect this and all the top trigrams from the previous months are replaced with references to the app. Another key difference present in the NHS accounts is the frequency of *thanks* and *thank you* phrases throughout the year. Although those phrases were present in WHO and CDC accounts, they were used at very low frequencies, shown in Figure 6.16. The NHS accounts use these phrases to thank individual users, nurses, doctors, healthcare workers, and the public for efforts related to COVID-19.

5. Keyword Analysis

Keyword analysis shows the "extent to which we can trust an observed frequency difference, irrespective of its corpus size" (Gabrielatos 2019). Keyness measures whether a word has a higher or lower frequency than a reference corpus; a "significant jump in the keyness of a word simply means that the word is more frequently used in a particular setting than was expected" (Kretzschmar et al. 2004: 40-1). Keyness results provide an additional perspective on the linguistic patterns and prevalent tokens, topics, and overall discourses shown through ngram analysis above, and they also provide further evidence for these differences in communication by

each organization. Keyword analysis is accomplished using R and R packages tidytext and quanteda (Silge & Robinson 2017, Benoit et al. 2018)[34]. Because the reference corpus has important implications for the keyword analysis results (Culpepper 2009: 43), each corpus is compared against the other two governmental organizations as a reference corpus to facilitate the most direct comparison possible. Keyword analysis has been shown to be a productive way to uncover both general genre differences and distinctions between texts of the same type. The top 100 keywords were manually inspected for token and category (shown for each unit in tables 6.4-6.6); then the top tokens and notable keywords are inspected through dispersion across 2020 and via concordance lines utilizing the kwic() function from polmineR (Blätte 2019).

CDC keyword analysis shows the words which index a variety of differences from the entire group. CDC keyword results are shown below.

Table 6.4: CDC Keywords

Semantic Group	**Keyword**	**Loglikelihood**
Individuals and People	your	358.46
	veterans	2423.14
	veteran	816.44
	provider	490.78
	adults	482.97
	American	458.33
	public	438.38
	providers	434.48
	doctor	265.97
Regions and Organization	CDC	5003.03
	program	1346.91
	state	453.36
	military	430.86
	states	406.36

[34] Matrix multiplication is conducted in R using contingency tables of the frequencies of each token compared to the reference corpora. In order to measure the differences of each governmental unit's communication against the others, the reference corpora differed slightly for each keyword analysis. The CDC accounts were measured against the WHO and NHS as a reference corpus. The WHO accounts utilize the NHS and CDC communication as a reference corpus, and the NHS keyword analysis utilizes the WHO and CDC as a reference corpus. Log-likelihood (G2), raw and normalized frequencies, the difference coefficient (Leech and Fallon 1992), and the relative frequency rations (RFR) (Damerau 1993) are all calculated for each keyword (Gries 2017: 197-200).

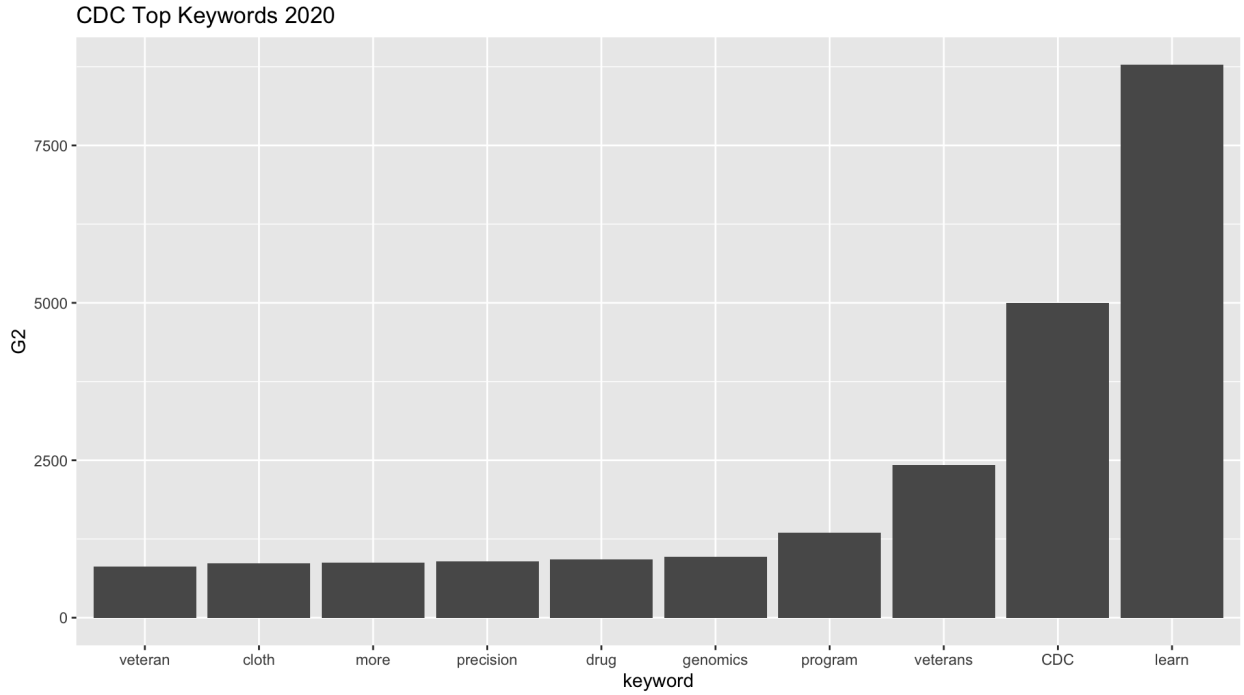
	programs	379.09
	federal	24.5
Medical-related tokens	genomics	964.53
	drug	927.58
	cloth	859.06
	flu	755.71
	sick	722.84
	#fluvaccine	556.73
	drugs	533.09
	illness	514.25
	healthy	460.78
	tests	400.29
	sanitizer	371.18
	disease	351.2
	cancer	342.3
	prescription	298.5
	#asthma	292.75
	hospitalization	280.7
	safety	276.3
	mask	273.8
	exposures	267.95
	pediatric	265.3
	shot	247.2
	#onehealth	225.75
Inanimate-objects		
	products	777.55
	food	389.25
	registry	362.11
	publications	266.56
	product	231
	chemicals	224.42
Animate-objects		
	pets	595.82
	pet	449.29
Characterizing attributes	precision	897.93
	more	876.23
	feet	775.8
	slow	744.05
	authorized	502.32
	risk	410.53
	certain	339.1
	molecular	339
	disaster	332.16
	environmental	318.49
	related	315.06
	higher	312.54

	heat	309.6
	defects	297.6
	hazards	297.2
	benefits	274.5
	authorization	266.89
	publications	266.56
	travel	253.6
	cause	248.1
	least	236.7
	generic	234.1
	new	233.9
Actions and Information	learn	8778.98
	about	787.82
	resources	755.71
	prevent	507.47
	coach	365.77
	use	363.39
	tips	348.78
	shows	298.85
	spread	292.2
	exposed	272.89
	study	265.6
	associated	262.14
	options	254.7
	travel	253.6
	approved	253.22
	among	252.4
	cause	248.08
	include	246.3
	finds	238.18
	how	237.5
	recommends	231.8
	help	230.72
	issued	223.8

Keywords for the CDC index a variety of unique aspects of their communication on Twitter. CDC keyword results in table 6.4 include semantic groups of actions and information, references to individuals and people, medical-related terms, inanimate and animate objects, and characterizing attributes. There are also frequent references to different types of illnesses and to medical terms: *sick, healthy, #fluvaccine, illness, disease,* and *cancer.* References to people include specific groups: *Americans, veterans, provider(s),* and *adults.* Characterizing attributes

include tokens like *certain, slow, molecular,* and *environmental.* Keywords *spread, learn, resources, associated,* and *approved* are top examples referencing actions and information.

Figure 6.17



The contexts of tokens with higher keyness scores were inspected via randomized concordance lines using R[35]. *Learn* has the highest keyness score for the CDC, and contexts of use include *learn more about how, read to learn, work with your doctor to learn,* and *watch to learn,* with some situations referring specifically to ways to learn more about COVID-19, like *join the CDC to learn the latest about the outbreak.* Additional situations of use reference other illnesses or illness prevention more generally. *Learn* is used in the context of pointing users and the audience(s) of these tweets to additional resources. An interesting concordance line discusses the US *working hard to learn more every day* about COVID-19 and outbreaks. Another frequent token with high keyness values is the *CDC* itself, which often occurs with the term *scientists,*

[35] Concordance lines and keywords-in-context are obtained via the kwic() function in polmineR (Blätte & Leonhardt 2019). Each kwic object is stored in a dataframe in R, which then allows for subsetting by corpus, year, and specific keyword searches. This object is then sorted using randomized sampling with the sample_n function on the rows of the concordance dataframe.

resources, or verbs like *recommends, collaborated,* and *remains committed.* The token *program* occurs with many official programs and initiatives of the CDC: *find a program near you this #NationalMammographyDay.* Next, the dispersion of top keywords was inspected through the dispersion function in polmineR (Blätte & Leonhardt 2019) to discover the consistency of usage across 2020. Figure 6.18 shows that *learn* was used at a similar frequency throughout the year, showing that the COVID-19 pandemic perhaps did not change this aspect of CDC communication via Twitter and the use of the token *learn* to describe additional resources or options for users.

Figure 6.18

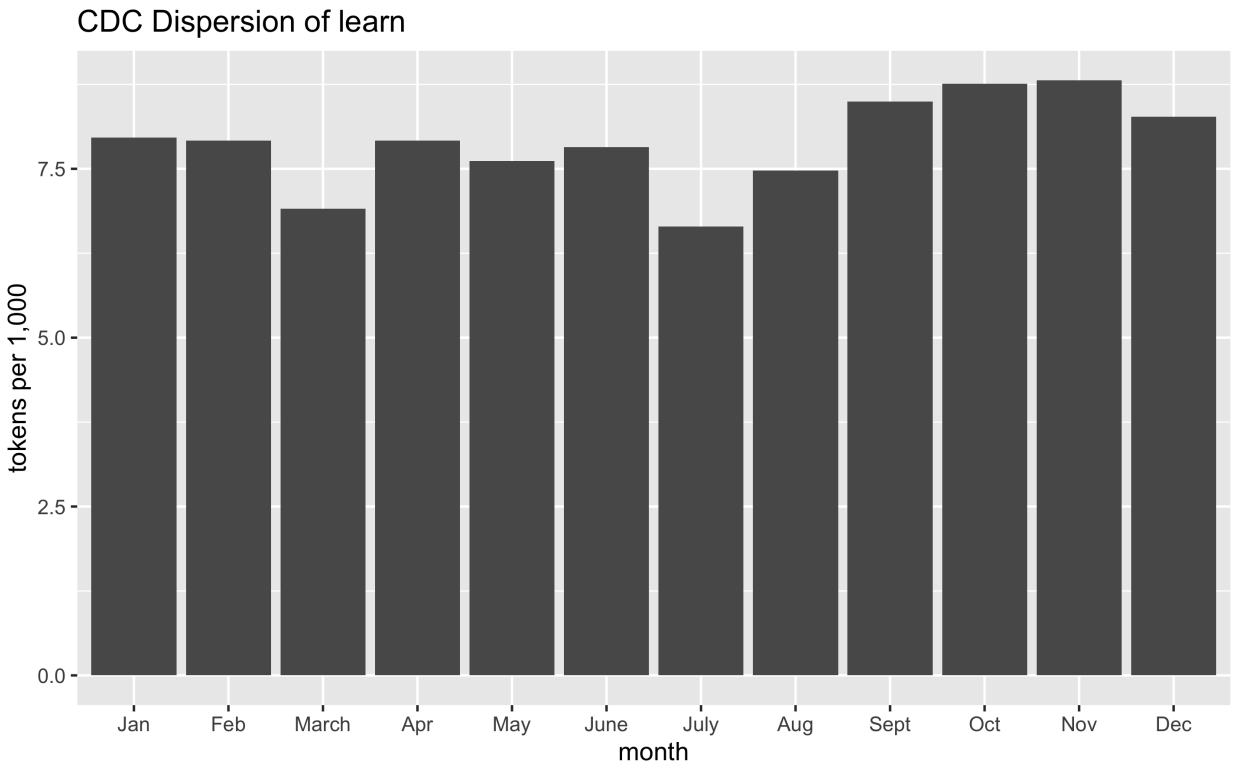


Table 6.5: WHO Keywords

Semantic Group	Keyword	Loglikelihood
Individuals and people	dr	912.83
	partners	458.9
	workers	384.99
	we	277.4

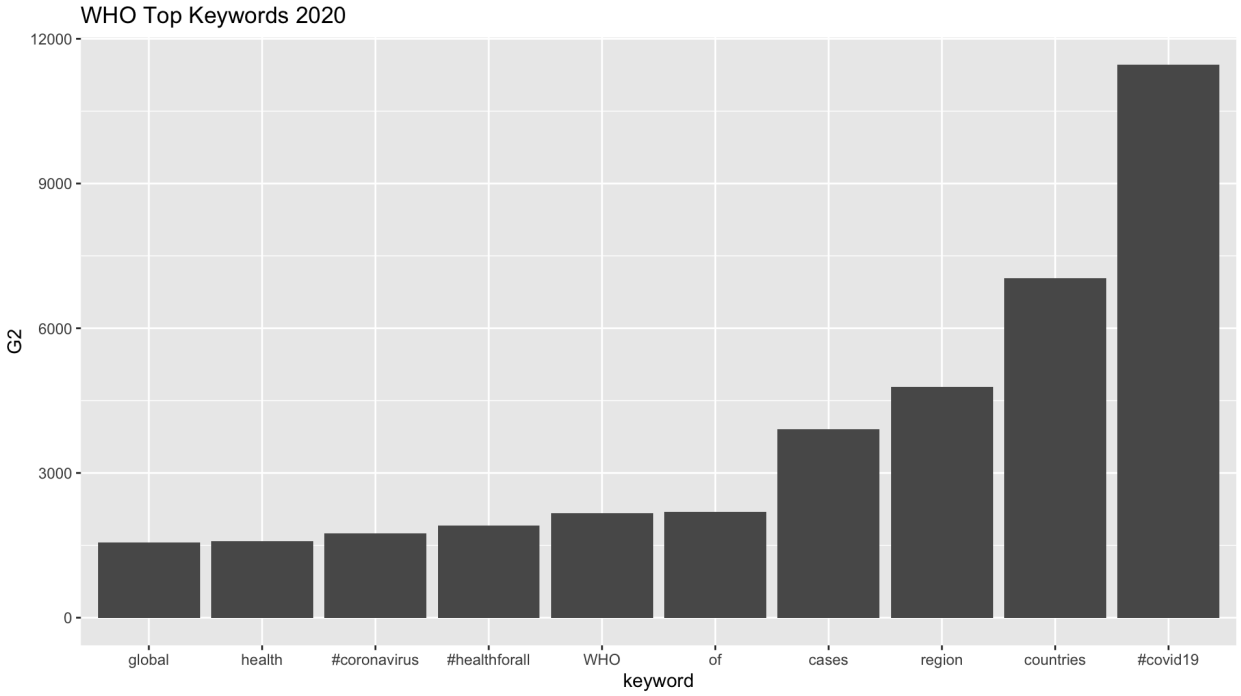
Regions and Organizations	WHO	2170.14
	global	1555.63
	African	1421.73
	world	1278.28
	countries	7042.79
	region	4780.12
	pacific	1098.47
	western	975.67
	#drc	963.87
	Europe	874.05
	Africa	766.84
	European	724.1
	continent	721.28
	international	621.7
	#africa	628.2
	country	616.95
	governments	454
	#china	442.5
	Americas	392.3
Medical-related tokens	#covid19	11459.14
	cases	3900.68
	#healthforall	1908.89
	health	1747.62
	#coronavirus	1589.60
	transmission	1074.6
	#ebola	782
	#supportnursesandmidwives	762.67
	#vaccineswork	645.01
	vaccines	587.78
	diseases	529.9
	immunization	405.3
	#mentalhealth	370.46
	#healthyathome	370
	polio	362
	ebola	352
	virus	332.5
	disease	312
	therapeutics	309
Inanimate-objects		
	tobacco	675.9
	supplies	498.6
	#notabacco	432
Animate-objects		
Information and Actions	response	1376.17
	pandemic	1270.46
	deaths	1166.37

	confirmed	1081.65
	million	989.75
	reported	846.81
	measures	686.55
	systems	665.55
	income	636.5
	briefing	597
	recoveries	542.9
	ensure	535.8
	must	522.45
	accelerator	447
	invest	415.2
	situation	384.6
	billion	369.2
	outbreak	352.7
	evidence	340.7
	economies	338
	figures	303
	capacity	292
Characterizing attributes	solidarity	1117.90
	regional	731.82
	economic	493.2
	essential	468.8
	universal	445.12
	preparedness	455.2
	violence	448
	globally	407.2
	equitable	351.78
	physical	328
	cumulatively	309
	political	304.6
	together	279

The WHO keywords reveal a variety of differences from the CDC and NHS accounts. Keywords in use by the WHO are predominantly references to regions and organizations like *WHO, global, countries, Europe,* and *pacific.* Information and actions include keywords *deaths, million,* and *pandemic*. Characterizing attributes reference economic indicators and qualities including *economic* and *equitable.* Very few references to individuals and people are present aside from several keywords: *#supportnursesandmidwives*, *suppliers, dr, workers,* and *we***.** The WHO keywords reference illness generally and specific diseases like *#ebola, Ebola, diseases,* and *polio.* Keywords also include names for COVID-19, including *#covid19* and *#ncov2019.* This

indicates that overall, the WHO's communication focuses on regions, case numbers, and deaths much more than the CDC or NHS. In addition, this also underscores findings in the trigrams, that there are few, if any recommended actions present in the WHO communication throughout 2020; instead they focus on sharing information about the spread of COVID-19 cases and regions where case numbers change.

Figure 6.19



Further inspection of concordance lines reveals particular aspects of the WHO's top keywords usage. The token *#COVID19* is used in the context of specific case counts and similar information: *Brazil surpassed one million #COVID19 cases, joining the United, challenges during #COVID19, minimize the risk of #COVID19.* At times it is also used with other names for COVID-19, including *#coronavirus* and *#COVID19. Countries* are referred to in reference to specific regions or locations: *Latin American countries, countries of the Americas,* and to COVID-19 case numbers: *as countries strengthen surveillance of #COVID-19, working with all countries*, and *as countries scrambled to respond to the pandemic. Cases* are not only describing

numbers for COVID-19 but also other types of diseases and health concerns: *explains that one in five cases of liver disease, majority of cancer cases worldwide, prevent severe cases of the flu.* Concordance lines reveal interesting uses of the token *health,* often being described in the abstract or generally: *health systems, health promotion, public health solutions, local health authorities, allow health officials to understand, avoid overwhelming the health care service.* These findings underscore previous examples from the trigram analysis and further results of inspection of keywords: the WHO tends to focus on information, health, and case numbers in the actual and the abstract, rather than on individuals and recommended actions for prevention and care. In order to further understand the contexts and frequency of use, dispersion is implemented on top keywords. The WHO accounts utilize the tokens *countries* and *health* to varying extents throughout 2020 (as shown in the figures below). *Countries* appears to be relatively stable across the year, with the largest change occurring from January to February, with a considerable jump in frequency of use. *Health* occurs with more variability throughout the year but is used in a wider variety of contexts and in reference to different health systems and as an abstract ideal.

Figure 6.20 WHO Countries Dispersion

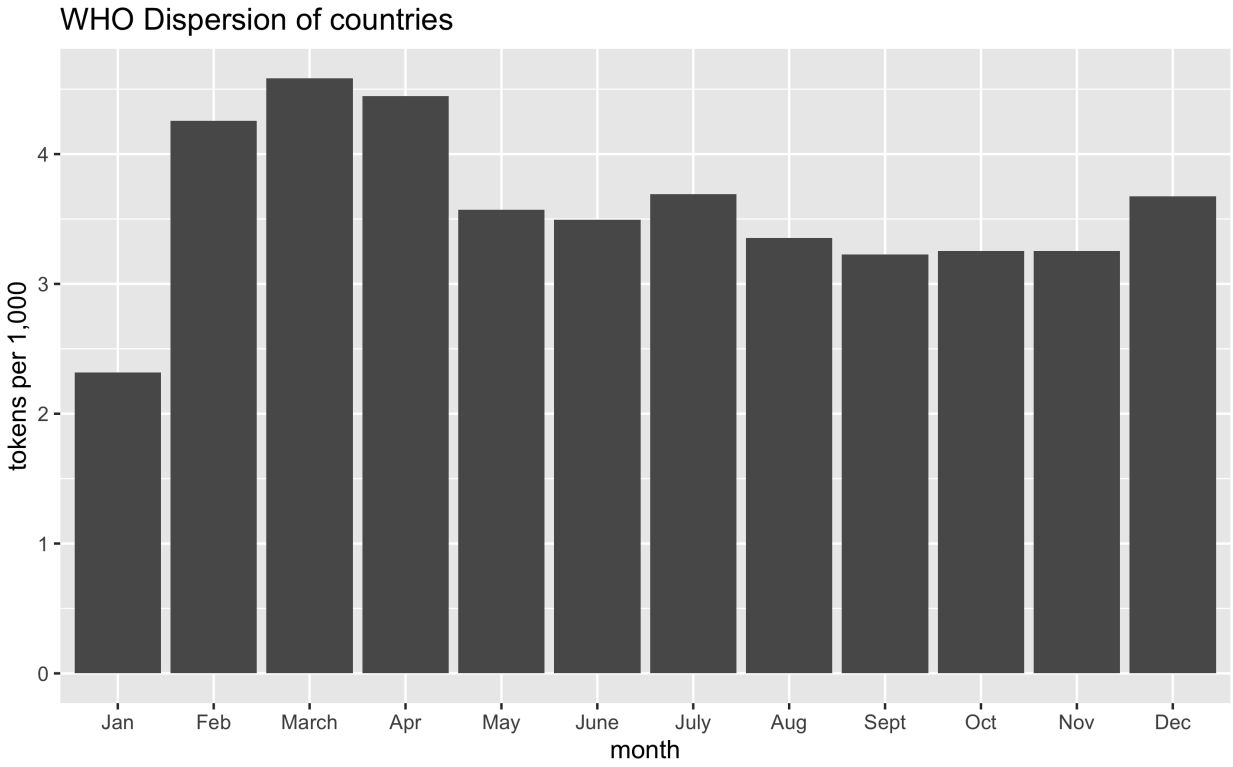


Figure 6.21 WHO Health Dispersion

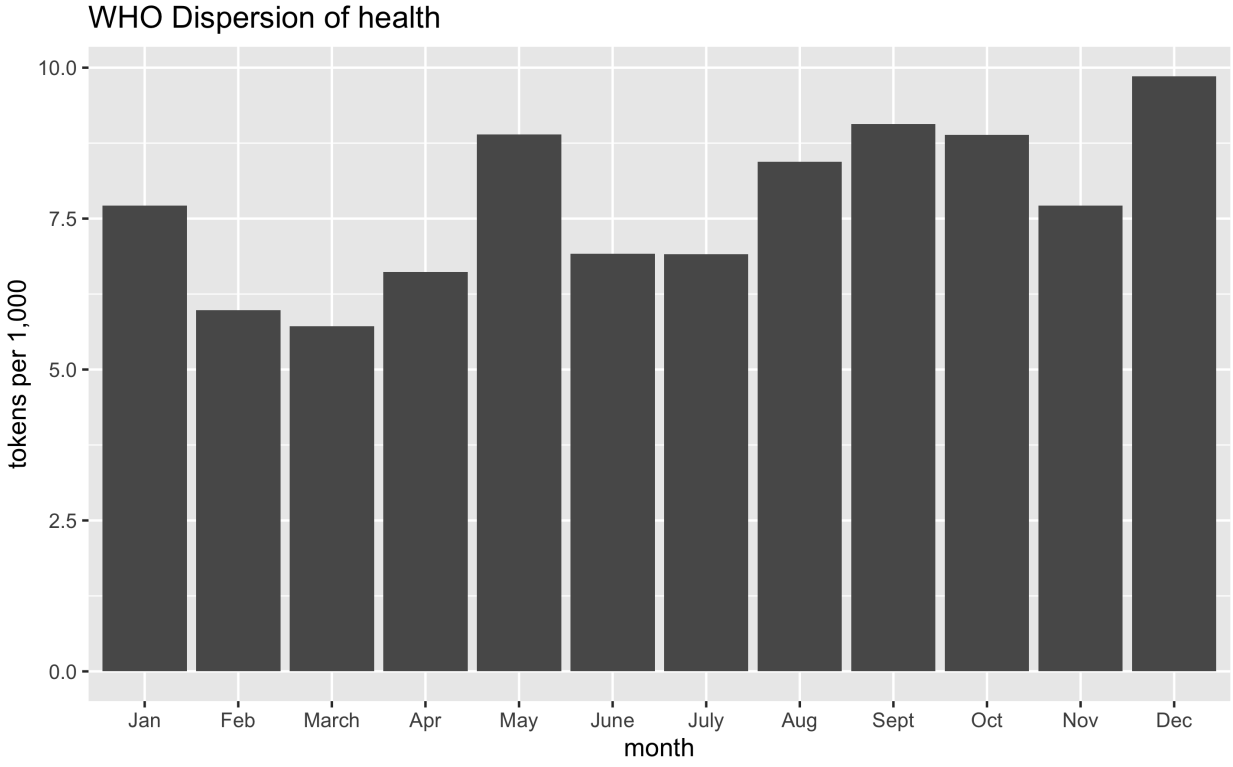


Table 6.6: NHS Keywords

Semantic Groups	Keyword	Loglikelihood
People and individuals	you	8412.97
	our	3136.6
	staff	1754.9
	self	1340.05
	you're	979.48
	minor	593.87
	team	591.5
	pharmacist	477.14
	carers	457.3
	colleagues	393.53
	we	391.05
Regions and Organizations	Wales	1209.42
	Scotland	1013.8
	Google	970.68
	#NHS	830.4
	England	820.53
	London	720.27
	helpline	552.11
	UK	463.7
Medical-related terms	jab	438.6
	advice	2768.8
	coronavirus	2228
	hospital	1529
	wellbeing	1139.59
	minor	593.87
	appointment	557.3
	dental	508.1
	symptoms	491.28
	maternity	426.07
	mental	401.28
	#hospital	393.6
Animate- objects		
Inanimate- objects		
	phone	706
	algorithm	701.6
	donation	631.85
	messages	467.1
	weekend	466.65
	venue	419.12
	alert	367.2
Information and Actions	find	2143.69
	visit	1429.14

	isolate	1105.08
	book	889.44
	support	797.2
	test	788.44
	hear	698.42
	alarms	671.8
	go	663.6
	referring	660.9
	sent	652.3
	automated	631.87
	look	624.6
	contact	609.32
	filter	585.2
	appreciate	570.74
	helpline	552.11
	asking	538.6
	need	536.28
	will	525.5
	dm	503.7
	visiting	501.7
	uses	472.95
	been	418.8
	link	417.7
	get	404.8
	feedback	393.7
	trace	362.7
Characterizing attributes	here	1393.3
	amazing	1043.5
	fantastic	427.24
	free	455
	chief	384.05
	urgent	372.9
	wonderful	368.9
	great	364.6
	positive	360
Greetings	hi	42328.45
	thanks	4251.9
	please	3433.7
	thank	1656.6
	sorry	933.46

The NHS results reveal many key differences from the governmental conversations and communication on Twitter. The keyword category greetings is only present in NHS accounts and includes keywords *hi, thank(s), please,* and *sorry.* This hallmarks a personalized response, and when taken into account with the full NHS communication, varied recommended actions and ideas are presented. They also utilize many different tokens and semantic categories, presented in

the top 100 keywords above, including references to individuals and people, medical-related terms, objects, characterizing attributes, information, and greetings. This is also highlighted through the variety of actions, like *isolate, test, support,* and *find.* The NHS also refers to many different groups of people and to individuals, including health professionals and members of the wider public like *staff, pharmacist, our,* and *you*, paralleled in key abstract concepts discussed like *advice, wellbeing, mental,* and *symptoms.* Taken as a whole these indicate comprehensive modes of support and recommended actions provided by the NHS via Twitter.

Figure 6.22

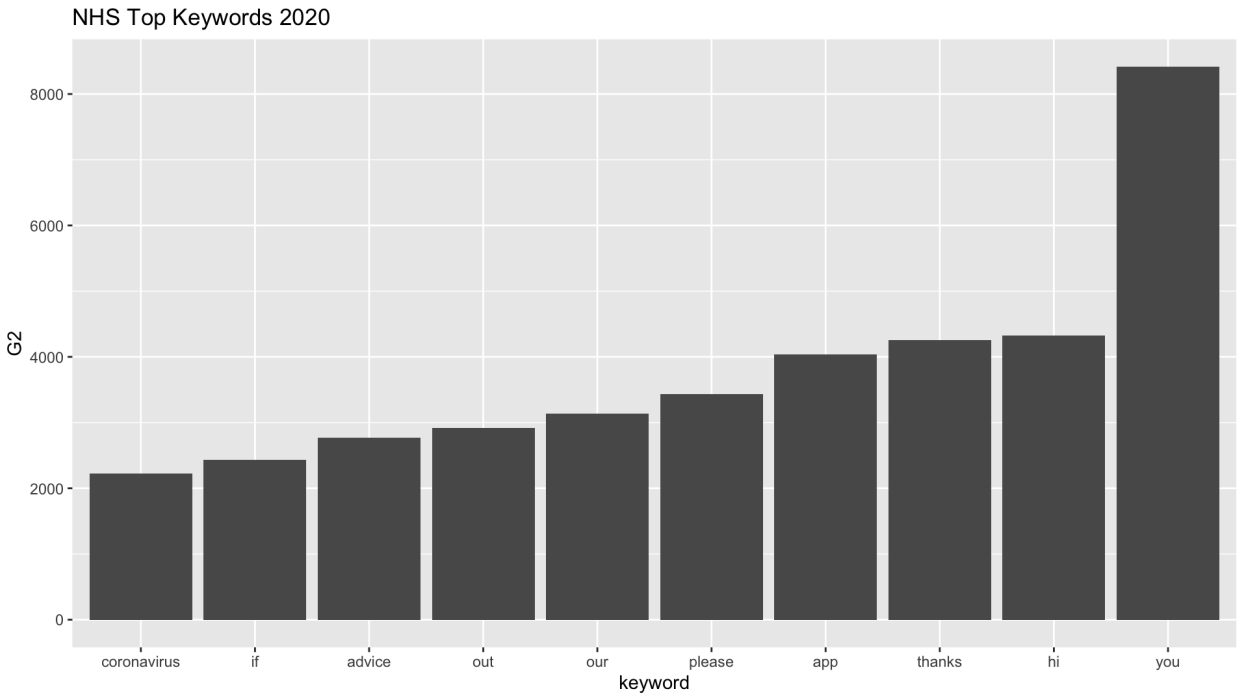
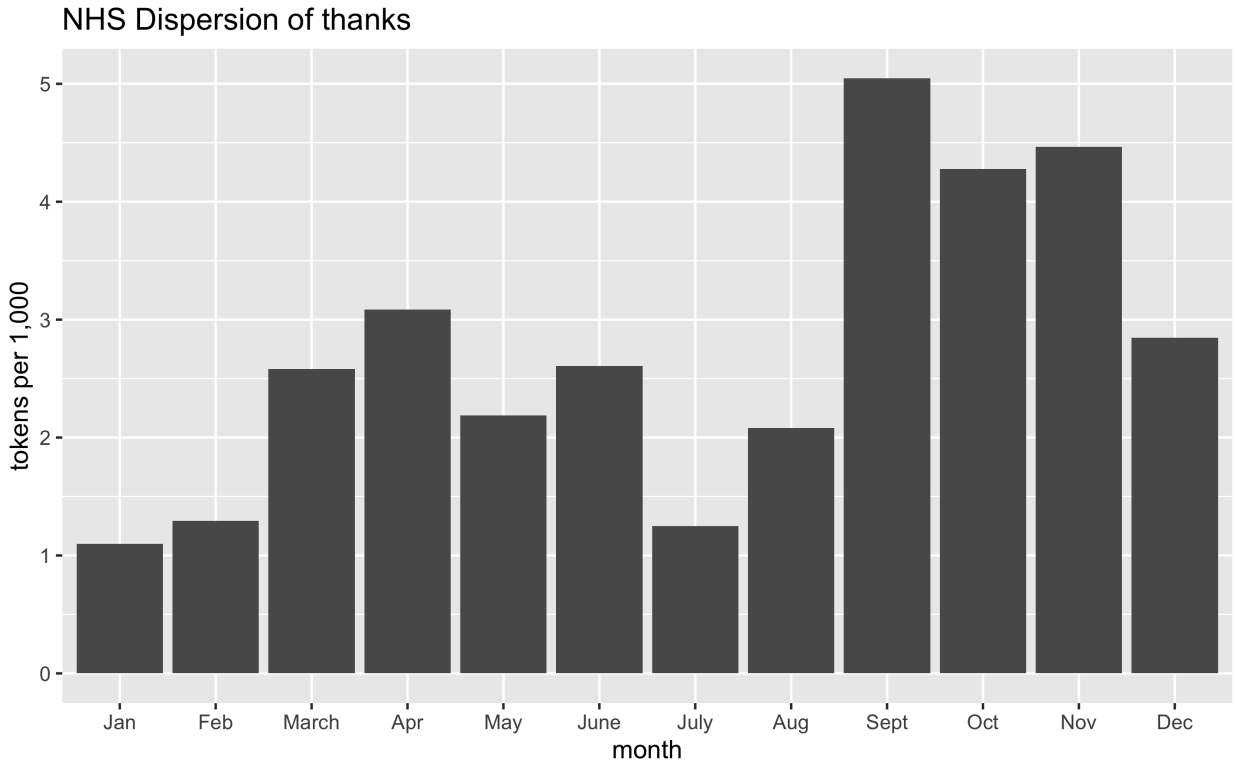


Figure 6.22 shows the top keywords in use by the NHS; contexts of usage reveal interesting discourses associated with each token. *Advice* is used in many different contexts and in reference to different issues: *for expert advice, tips and advice, self-care advice, helpful, free advice,* and *advice and reassurance for cancer patients.* It is also used specifically in reference to the COVID-19 pandemic: *following the latest advice, online service for quick health advice.* Tokens *you* and *thank* are often used together; *saying thank you to everyone, thank you for your*

feedback, and *we want to say thank you to each and every one* are some examples of contexts of use. *Thank you* is also used to thank specific people and individuals such as this tweet: *thank you Ava.* The term *you* is also utilized in situations inviting individual users and the directed audience of these tweets to take advantage of NHS offerings, help, and knowledge: *have you been trying to call our, alternatively you can email,* and *if you have symptoms of cancer that you are worried about. Thanks* and *thank* are very interesting keywords that are not characteristic of CDC and WHO communication. Figure 6.23 illustrates the variability in use of this token in NHS communication. *Thanks* is occurring most frequently beginning in September 2020 and peaks in use in the spring in April of 2020. This makes sense considering circumstances both worldwide and in the UK specifically.

Figure 6.23



Many keywords are a result of this communication being on Twitter, with hashtags added to single tokens or phrases, often for emphasis. Hashtags are also routinely applied by each

organization to different initiatives or programs created by each organization; for example, the NHS test and trace program includes a variety of hashtag phrases used on Twitter to discuss and spread the word about the program like #nhscovid19app**.** The WHO accounts include the most keywords with hashtags like *#healthyathome, #supportnursesandmidwives, #healthforall,* and *#ebola*. The CDC also applies hashtags to specific health problems: *#flu, #asthma* and to initiatives: *#onehealth.*

6. Collocations

The following section discusses results of collocational analysis[36] of specific tokens, including *COVID-19, #COVID-19, coronavirus*, and top keywords of each organization. Top collocations for names of COVID-19 underscores findings from each previous type of analysis.

Table 6.7: COVID-19 Top Collocations

CDC	**WHO**	**NHS**
spread (858)	cases (1271)	for (922)
more (738)	with (1171)	have (787)
during (549)	pandemic (1141)	you (779)
learn (524)	#coronavirus (942)	NHS (486)
you (471)	during (842)	our (455)
pandemic (444)	WHO (759)	is (451)
help (415)	confirmed (691)	app (443)
your (413)	response (630)	coronavirus (438)
is (380)	health (606)	during (411)
have (368)	more (536)	has (410)
protect (352)	African (506)	from (397)
slow (332)	#covid19 (500)	positive (356)
health (309)	region (453)	been (344)

Top collocations for variant spellings of COVID-19 emphasize the differences in patterns for each organization; CDC top collocations focus on actions related to COVID, including *spread, learn, help, protect*, and *slow.* Other CDC collocations reflect a connection to individuals and health largely: *you, your, help,* and *health.* WHO top collocations primarily are connected to information, regions, and cases and the World Health Organization itself: *cases, pandemic,*

[36] Collocations are calculated utilizing the cooccurrences() function in polmineR (Blätte & Leonhardt 2019); they are ranked in order by log-likelihood (Dunning 1993) and raw frequencies.

confirmed, we, and *WHO, response.* WHO collocations also include various names for COVID like *#coronavirus* and *#covid19.* Collocations for COVID in NHS accounts reference the NHS itself, individuals, and the COVID-19 app: *you, NHS, app, coronavirus, during, positive.*

Table 6.8: Coronavirus Top Collocations

CDC	**WHO**	**NHS**
of (178)	#covid19 (919)	for (922)
novel (80)	on (568)	have (787)
spread (75)	drtedros (513)	you (779)
in (68)	cases (452)	NHS (486)
disease (67)	more (431)	our (455)
#covid19 (55)	region (309)	is (451)
2019 (55)	info (298)	app (443)
is (52)	new (277)	coronavirus (438)
new (45)	@drtedros (217)	during (411)
force (40)	WHO (165)	has (410)
task (40)	➡️ (159)	from (397)
prevent (39)	are (146)	positive (356)

Collocations for *coronavirus* and *#coronavirus* emphasize differences in frequencies of use across all three corpora. WHO is the main organization that continues to utilize variant spelling of *coronavirus* and other names for COVID-19 throughout the year. Collocations in the WHO accounts are primarily relevant to information: *cases, more, region, info,* and *new.* CDC collocations also include descriptive information and characterizing attributes of COVID-19, like *novel, new, 2019,* and *disease,* with other tokens reflecting actions relevant to the wider public and the CDC: *task, force,* and *prevent.* The NHS includes the most collocations relating to individuals and applicable information: *you, spread, symptoms, you, help,* and *positive.*

Table 6.9: CDC Collocations of Top Keywords

CDC	**Learn**	**program**
new (397)	more (7330)	for (183)
that (385)	to (3619)	health (176)
is (372)	about (3292)	more (133)
health (346)	how (2040)	learn (106)
of (329)	your (1154)	VA (103)
with (291)	you (933)	about (102)
more (258)	for (797)	national (93)
read (233)	can (774)	veterans (93)
on (230)	health (709)	you (88)

recommends (228)	this (662)	can (85)
latest (213)	at (619)	help (83)
learn (209)	is (506)	CDC's (81)
report (207)	from (499)	your (65)
has (200)	help (244)	medical (56)

Top keywords are also analyzed for each organization; the top keywords analyzed for the CDC include *CDC*, *learn*, and *program*. Collocations of top keywords are broadly reflective of the priorities in the CDC's communication throughout the year. Characterizing attributes make up the majority of top collocates for the token *CDC* including *more, new, latest,* with additional collocates referring to actions like *learn, report, read,* and *commands.* The top keyword *learn* includes collocates connecting to individuals like *you, your, health* and places to access additional information. *Learn* is one of the very frequent tokens in the CDC corpus, a prevalent token utilized in 2020. *Program* is used as a general term to connect to broader initiatives and CDC schemes for the public. Collocations of *program* reflect a US-internal focus in these initiatives and goals of promoting health: *help, CDC, VA, medical,* and *national.*

Table 6.10: WHO Collocations of Top Keywords

WHO	**cases**	**regions**
to (1532)	in (1518)	in (129)
in (928)	of (1287)	other (84)
with (856)	#covid19 (1203)	countries (67)
is (775)	confirmed (813)	WHO (66)
region (523)	reported (621)	daily (56)
#covid19 (504)	on (561)	follow (46)
have (504)	more (493)	for (44)
people (456)	deaths (492)	all (26)
has (431)	#coronavirus (441)	to (25)
health (404)	for (437)	#coronavirus (23)
more (377)	region (422)	check (20)
Europe (295)	have (304)	data (20)
that (277)	over (288)	#covid19 (13)
from (264)	info (286)	Is (11)

WHO collocations emphasize a focus on COVID and on information relevant to case numbers around the world. Broad patterns are reflective of this emphasis and on the focus on locations and regions throughout the pandemic. Collocations of all three top keywords *WHO,*

cases, and *regions* reflect these themes. The top keyword *WHO* collocates most often with *region, #covid19, people, health,* and *Europe.* Collocational patterns with keywords *cases* and *regions* include the tokens *confirmed, reported, deaths, #covid19, #coronavirus, region, info, countries, daily,* and *data.*

Table 6.11: NHS Collocations of Top Keywords

you	**thanks**	**hi**
if (9764)	for (1822)	thanks (1642)
can (7873)	hi (1642)	you (851)
your (7127)	your (826)	to (674)
for (6012)	you (622)	we (496)
are (5442)	here (597)	app (490)
have (5395)	more (458)	sorry (481)
you (5312)	our (440)	for (401)
need (4264)	out (398)	if (378)
thank (4116)	us (394)	more (326)
help (3807)	this (371)	are (304)
is (2681)	Debbie (355)	Debra (255)
with (2679)	Debra (341)	can (250)
do (2557)	we (264)	Debbie (242)
be (2544)	all (243)	your (237)
know (2313)	so (233)	hear (234)
this (2098)	much (229)	have (206)
get (2012)	problem (200)	here (205)
find (1943)	❤️ (194)	please (190)
NHS (1877)	Emma (190)	Emma (171)
we (1852)	more (458)	we’re (156)
here (1776)	our (440)	Pamela (132)

NHS collocations of top keywords further emphasizes the distinctive and diverse nature of their communication efforts through Twitter. Collocations of the token *you* include additional references to individuals: *your, we* and to recommendations: *help, get, find, need.* Collocates *thank* and *here* illustrate the efforts to give individuals as many opportunities possible for addressing diverse concerns. The collocational patterns with *hi* and *thanks* reflect this individualized approach, with many collocations including specific individuals and proper names like *Debbie, Debra, Emma,* and additional greetings present like *sorry, please,* and pronouns *you, your, us,* and *we.*

The keywords and their collocations discussed above reveal and underscore differences highlighted by the trigrams; they provide important evidence in the story of communication throughout the year. The CDC accounts include reference to different options for individuals, to many different diseases, and to Americans more broadly. The WHO accounts differ from the NHS and CDC in that they focus overwhelmingly on worldwide information in the form of case numbers and regions and at times discuss health in both general and abstract concepts like *health systems* and *public health.* The NHS stands apart in that they include comprehensive variety in offerings and options for users, tailored to individuals, sharing wide-ranging facts and information.

7. Conclusion

This is a crucial dataset for a number of reasons. "Rumors thrive in situations of uncertainty whether around a new vaccine, unfamiliar disease outbreak, or more catastrophic events like wars, natural disasters, or pandemics" (Larson 2020: xiv). Public health messaging, like all media and government communication is subjective and curated. Trigram analysis provides linguistic evidence of the topics, priorities and methods used to communicate about these topics, in this specialized dataset. Keyword analysis highlights the major differences across each organization and provides insight into changes over time during the year in each group's use of keywords[37]. Given that the NHS as an organization involves many health and medical practices, the variety in their messaging is perhaps unsurprising. The WHO accounts differ from the NHS and the CDC inclusion of few, if any references to individuals and little to no frequently used trigrams with advisory phrases. Despite the dramatic and harrowing developments throughout 2020, stable trigrams were present across all three institutions, and these trigrams played important functions in terms of organizing discourse, introducing topics, and referential functions. *The spread of, find out more, to learn more,* and *the #covid19 pandemic* are a few examples.

[37] Further exploration of these keywords and their dispersion is warranted.

Using these methods together enables a detailed view of these governmental agencies' messaging throughout 2020. Trigrams reveal dominant patterns, with stability and variation in messaging choices by the CDC, WHO, and NHS accounts. Keywords are especially useful for emphasizing the differences in communication across each agency and provide evidence for the associated and related discourses. Collocational analysis and dispersion reveal the dominant patterns associated with keywords in use across agencies, in addition to fluctuations and trends in messaging.

This chapter illustrates the merit and advantages of corpus linguistics applied to health and crisis communication and to social media. Baker writes that

> language change, perhaps particularly lexical change, has the potential to tell us much about societal change. Language does not develop in isolation but has a dialectical relationship with culture, both reflecting and spurring on changes in everyday life (2011: 2).

These corpora are a very specific and specialized genre of language use, in terms of the target audiences, the authors of these texts, and the broader language situation. Public health communication, especially from officials, is encompassed in the goals of both informational and advisory statements so that individuals will be empowered to make informed decisions (Regidor et al. 2007: 93). What happens when the available scientific information is changing or rapidly developing? Further questions remain including whether the present goals and imperatives of each organization align with the linguistic patterns and broader themes present in these discourses, data, and what recommendations might be for enhancing their efforts. Additional questions concern differences in messaging by the distinctive accounts within each organization and whether the accounts provide unified or fractured information and advice, as well as the relationship between tweet authorship with linguistic patterns and choices. Continued analysis and interrogation of public health messaging, especially by such critical units as the CDC, WHO, and NHS is also warranted for supporting efforts to mitigate crises like the COVID-19 pandemic.

CHAPTER 7

CONCLUSION

This work focuses on health communication in English across different genres and contexts. Each chapter of analysis utilizes corpora from diverse time periods and locations including the first English language scientific journal, US and UK press, and present-day governmental communication via Twitter. This research argues that corpus-based linguistics provides an advantageous and key contribution to existing studies of health and scientific communication. Corpus methods underscore previous findings highlighting the integral nature of linguistic patterns and choices, especially in public health and scientific communication. These methods reveal fundamental and competing narratives and discourses related to health, science, and risk in a variety of trends. There are clear findings in each chapter, made available using corpus-based methods combined with careful interpretation. These quantitative methods involve the use of statistical measures and frequency analyses[38] interpreted through core perspectives about language in use. The data and output along with these perspectives provide the necessary evidence for the story of public health language in each of these notable contexts.

The theoretical foundation for this research includes several crucial perspectives, studies of language as a complex system (Kretzschmar 2009, 2015, 2018) and work by British corpus linguists like Firth (1935, 1967), Sinclair (2004), Stubbs (2001, 2002), and Baker (2006, 2010, 2011, 2014). These perspectives emphasize the importance of studying language in use, the ongoing process of an emergent complex system. Properties of a complex system include continuous "dynamic activity", random interaction, exchange of information, reinforcement, and the "emergence of stable patterns" (Kretzschmar 2015: 11). Complexity science enables

[38] These methods all involved the use of CQP and R; see Chapter Three for a more detailed description.

researchers to take into account all the intricate factors involved in linguistic processes, including the interaction between cultural discourses and linguistic patterns (Kretzschmar 2015).

Each chapter shows that the predictions involved in complexity science are fulfilled. The patterns of linguistic variation always follow an asymptotic distribution, exemplified in each method and type of pattern studied, including frequency lists, collocations, and keywords. Stubbs and Sinclair argue similarly that the context of language use must be taken into account when studying linguistic variation, as it "involves both routine and creation…[and] transmits culture" (Stubbs 1996: 23-4).

Foregrounding the historical, linguistic, and cultural background is a crucial aspect of this work. Berridge and Loughlin point out that research on public health policy often tends to presuppose that historical events are "distant…that the history is 'background' which stops some while before the present day" (2004: 2). Instead, the cultural influences including historical events and competing narratives, are rooted integrally in public health messaging. When considering the effects of cultural discourses and historical events on linguistic patterns, "the transfer and transformation of ideas, organizational innovations and technologies internationally" (Haines 2005: xvi) have all played a major role in terms of health communication.

These methods alongside the theoretical background facilitates a nuanced and detailed perspective of public health language and rhetoric, using computational methods by integrating CQP with R packages like polmineR (Blätte & Leonhardt 2019). These methods and their corresponding output are useful, necessary tools to provide the evidence and basis for conclusions and interpretation of the linguistic patterns within. Another core aspect of these methods involves careful corpus sampling and compilation, underscored by previous work including Kretzschmar et al. 2004. This enables productive evidence-based output so that effective interpretation and conclusions may be drawn in a meaningful manner.

Each chapter demonstrates integral findings on the different data sets and language situations, along with broader conclusions and applications of this work. The corpus-based

methods also emphasize the discourses associated with the data trends, in addition to exposing the actual linguistic patterns in use by context. Key methods include frequency analysis, keyword analysis, and collocational analysis; different variations of these are present throughout this research, displaying the range of possibilities and utility of corpus linguistics. The flexibility and utility of these methods incorporates different options depending on the research questions and goals.

The study of the *Phil Trans* demonstrates the effects of changing scientific progress and increased knowledge on linguistic patterns and on medical practices. The *Phil Trans* is the largest dataset in this work and includes 400 years of advancing medical and scientific views. The keywords *health, disease(s),* and *inoculation* under scrutiny exemplify core findings involving changing linguistic patterns and discourses. The *Phil Trans* authors shift from Early Modern humoral theories to modern scientific-based practices and a broader focus on public health culminating in the nineteenth century.

Chapter Five on US and UK press highlights the impact of trends in media portrayals of e-cigarettes, with predominant differences across locations. This chapter focuses on the variation in collocational patterns and bigrams over time across sources. The US generally emphasizes business and economic concerns, along with a greater tendency to focus on younger users over time and US-internal concerns. The UK tends to focus on worldwide concerns related to e-cigarettes, people in general, and cites outside research. These findings emphasize the complicated and layered nature of health-technologies and competing perceptions, with no overwhelmingly negative or positive depictions of this contested technology.

The last chapter of analysis provides another important scenario in public health communication, underscoring another venue of the importance of the linguistic patterns when providing health-related messaging. The CDC, WHO, and NHS have each played a prominent role in public health worldwide and in navigating the COVID-19 pandemic, with a small part of their efforts involving communicating via Twitter. The keywords, collocations, and ngrams used

by each group reveal distinctive communicative choices, which also expose their priorities and associated discourses. The CDC provides different types of information with more recommended actions over the course of 2020, while the WHO emphasizes case counts, regions, and COVID-19 itself. Each agency should take further action in diversifying their health messaging on Twitter to mitigate ongoing crises related to COVID-19, including providing additional resources and recommendations for individuals and communities. The World Health Organization's messaging would be enhanced by specifying more advice and recommendations for individuals, rather than just information about cases. The CDC should also focus on delivering resources and advice in a multitude of ways to reach a wider audience. The NHS includes the greatest variation in communication, referring to individuals and groups of people, providing information, and different types of recommended actions. The NHS also refers directly to COVID-19 the least out of all three organizations. At the end of 2020, NHS messaging primarily focused on the NHS COVID-19 application (app), rather than continuing to provide as many recommendations and resources as possible. Their Twitter accounts would certainly benefit by returning to the diverse messaging tactics overall, in addition to relevant information about the app.

Core takeaways from this work overall include the diversity of linguistic patterns when describing health-related topics across time, space, and genre. No distributional pattern or frequency is the same for any token, multiword unit, keyword, or collocation at any point in time. Public health language is contested, fraught with issues in reception and understandings of use by professionals, media coverage, and the public. Some examples of this persist throughout each chapter, including the patterns surrounding inoculation in the RSC, descriptions of health and e-cigarettes in US and UK press, and messaging about COVID-19 on Twitter.

Public health language also reflects cultural discourses which include sometimes negative, positive, or polarizing views. Interestingly in the case of the US and UK corpora on e-cigarette press, media coverage remains neither predominantly negative nor positive. The role of mass media in public health has only expanded over time (Berridge & Loughlin 2004: 6) making

this study a core example of media impact in distributing health information. “Medico-scientific and health news is now part of a process of production and dissemination that can have enormous and reciprocal policy impacts” (Berridge & Loughlin 2004: 6). News outlets and the media more largely “has been enlisted as a public health tool” (Berridge &Loughlin 2004: 6).

When examining changes in health-related communication over time, the contexts largely reflect genre-norms of each communicative situation. “Each genre of text” operates with important “social and cultural functions” (Stubbs 1996: 9). The *Phil Trans* includes the highly frequent use of first person and personal pronouns and other norms of the epistolary until scientific reporting, increased passivization, and new technical vocabulary emerged in the burgeoning and modernizing scientific disciplines. US and UK press follow expected norms for news articles, including specific emphasis on certain topics related to e-cigarettes. Twitter accounts also follow genre expectations, with each governmental organization utilizing brief, short posts and messages and sometimes including emojis or images in their Tweets.

By analyzing these distinctive situations and linguistic patterns, several central themes emerge, in addition to further questions and areas of research. Issues in power and trust pervade these communicative contexts in distinctive and fundamental ways, historically and presently. Disputes surrounding vaccination and inoculation are a notable example, and “attitudes towards vaccination on the part of the public remain…wary, compromised by uncertain science and decreasing awareness of the diseases vaccination prevented” (Berridge 2016: 104). The tobacco industry provides an important example of this, with major influences historically impacting present communication and key issues related to e-cigarettes. Problems with communicating about health with understanding and with empathy highlight these situations of power like governmental messaging on Twitter. A further relevant question includes the effects related to communicating from a position of power and the dynamics of risk communication broadly.

This work also highlights the importance of social networks, networks of communication, and professional networks, along with the prevalence of echo-chambers whether related to public

mistrust of vaccines, government communication, or scientific networks and confirmation bias. The interaction of competing cultural discourses and narratives, including what is scientifically valid, discussions on personal rights, and related concerns are also prevalent throughout this work.

There is a wealth of future research for linguists and public health studies related to this present endeavor. Some key questions include: what linguistic patterns used in communicating science remain predominant into the present day and their effects on both professional and public understandings. In addition to this, persistent parallels in historical health communication and in our current world are also worth further consideration.

The effects of governmental communication efforts on matters of grave public health concern like the COVID-19 pandemic and the effects of specific linguistic choices in messaging in each situation also necessitate further consideration. What can linguists do to better promote and enable productive communication in these spheres so that individuals, media coverage, and professionals can all be fully equipped to make evidence-based decisions?

This work underscores importance of the linguistic patterns and lexical choices when communicating about health and risk. “What counts as public health and what is emphasized will continue to vary depending on external factors…and which institutions and groups are involved” (Berridge 2016: 105). Regardless of what counts, the linguistic patterns involved are integral with tangible impacts historically, present, and in the future. It is widely applicable to real-world issues and concerns, including the COVID-19 pandemic and any situation of health or science communication, making continued research and investigation relevant and essential.

REFERENCES

Ackernecht, Erwin H. 1982. *A Short History of Medicine.* Baltimore: Johns Hopkins University Press.

Adelman, Joseph M. 2019. *Revolutionary Networks: The Business and Politics of Printing the News, 1763-1789.* Baltimore: Johns Hopkins University Press.

Adolphs, S., Brown, B., Carter, R., Crawford, P. ND. Clinical Linguistics: Corpus Linguistics in Health Care Settings. Center for Health Language Research (CHLR).

Adolphs, S., Brown, B., Carter, R., Crawford, C., and Sahota, O. 2004. Applying Corpus Linguistics in a health care context. *Journal of Applied Linguistics.* 9-28.

Adolphs, S., Atkins, S., Harvey, K. (forthcoming). Caught between professional requirements and interpersonal needs: vague language in healthcare contexts. *Vague Language Explored,* ed. By J. Cutting, Basingstoke: Palgrave.

Aratani, Lauren. 2020. How did face masks become a political issue in America? *The Guardian online.* *https://www.theguardian.com/world/2020/jun/29/face-masks-us-politics-coronavirus*

Arnold, Taylor B. 2020. cleanNLP package: A Tidy Data Model for Natural Language Processing. V. 3.0.3. Online: https://statsmaths.github.io/cleanNLP/

Atkins, Sarah, and Kevin Harvey. 2010. How to use corpus linguistics in the study of health communication. In *Routledge Handbook of Corpus Linguistics*, ed. by Anne O'Keefe and Michael McCarthy. 605-619.

Atkinson, Dwight. 1999. *Scientific Discourse in Sociohistorical Context: The Philosophical Transactions of the Royal Society of London, 1675-1975.* Lawrence Erlbaum Associates, Publishers.

Atkinson, Dwight, and Ellen Valle. 2013. Corpus Analysis of Scientific and Medical Writing Across Time. *The Encyclopedia of Applied Linguistics,* ed. by Carol Chapelle. Blackwell Publishing.

Baker, Paul. 2006. *Using Corpora in Discourse Analysis.* London: Continuum.

Baker, Paul. 2010. *Sociolinguistics and Corpus Linguistics.* Edinburgh: Edinburgh University Press.

Baker, Paul. 2011. Times May Change, But We Will Always Have Money: Diachronic Variation in Recent British English. *Journal of English Linguistics.* SAGE Publishing.

Baker, Paul, Costas Gabrielatos, and Tony McEnery. 2013. *Discourse analysis and Media Attitudes: The Representation of Islam in the British Press.* Cambridge University Press.

Baker, Paul. 2014. *Using Corpora to Analyze Gender.* Bloomsbury Publishing.

Baker, Paul, Gavin Brookes, and Craig Evans. 2019. *The Language of Patient Feedback: A Corpus Linguistic Study of Online Health Communication.* London, UK: Routledge Publishing.

Banda, Juan M., Tekumalla, Ramya, Wang, Guanyu; Yu, Jingyuan; Liu, Tuo; Ding, Yuning; Artemova, Katya; Tutubalina, Elena; Chowell, Gerardo. 2021. A large-scale COVID-19 Twitter chatter dataset for open scientific research - an international collaboration.

Baron, A. and Andrew Hardie. 2012. Prerequisites to a corpus-based analysis of EEBO-TCP. Presented at the EEBO-TCP 2012 conference, Oxford, United Kingdom. 17-18 September 2012.

Battersby, John. 2006. Translating policy into indicators and targets. *Oxford Handbook of Public Health Practice: 2nd Edition,* ed. by David Pencheon, Charles Guest, David Melzer, and J.A. Muir Gray. Oxford, UK: Oxford University Press.

Bednarek, Monika. 2016. Investigating evaluation and news values in news items that are shared through social media. *Corpora* 11(2), 227-257. Edinburgh: Edinburgh University Press.

Benoit, Kenneth, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, Akitaka Matsuo, William Lowe. 2018. quanteda: An R package for the quantitative

analysis of textual data. *Journal of Open Source Software*, 3(30), 774. Online: https://quanteda.io.
Bernhardt, Jay M. 2004. Communication at the Core of Effective Public Health. *American Journal of Public Health* 94(12).
Berridge, Virginia, and Kelly Loughlin. 2004. *Medicine, the Market and the Mass Media: Producing Health in the Twentieth Century. Routledge Studies in the Social History of Medicine.* Taylor & Francis.
Berridge, Virginia. 2007. *Marketing Health. Smoking and the discourse of public health in Britain, 1945-2000.* Oxford University Press.
Berridge, Virginia, Martin Gorsky, and Alex Mold. 2011. *Public Health in History. Understanding Public Health,* ed. by Ros Plowman and Nicki Thorogood. McGraw Hill: Open University Press.
Berridge, Virginia. 2014. Electronic cigarettes and history. *Lancet.*
Berridge, Virginia. 2016. *Public Health: A Very Short Introduction.* Oxford, UK: *Oxford University Press.*
Biber, Douglas, and Edward Finegan. 1997. Diachronic relations among speech-based and written registers in English. *To explain the present: Studies in the changing English language in honour of Matti Rissanen,* ed. by Tertu Nevalainen and Leena Kahlas-Tarkka, 253-275. Helskini: Société Néophilogique.
Biber, Douglas, S. Johansson, G. Leech, S, Conrad, and E. Finegan. 1999. *The Longman Grammar of Spoken and Written English.* Harlow, England.
Biber, Douglas, Susan Conrad, and V. Cortes. 2004. If You Look at…Lexical Bundles in University Teaching and Textbooks. *Applied Linguistics,* 371-405.
Biber, Douglas, and Bibieri, Federica. 2007. Lexical bundles in university spoken and written registers. *English for Specific Purposes.*
Biber, Douglas, and Susan Conrad. 2019. *Register, Genre, and Style: 2nd Edition.* Cambridge, UK: Cambridge University Press.
BIT Report. 2011. Behavioural Insights Team: Annual update 2010-11. Casaa.org.
Blätte, Andreas, Bernard Desgraupes, Sylvain Louiseau, Oliver Christ, Bruno Maximilian Schulze, Stefan Evert, Arne Fitschen. 2019. RccpCWB package, v 0.2.8. Online: https://www.github.com/PolMine/RcppCWB
Blätte, Andreas, Christoph Leonhardt. 2019. PolmineR() package, v 0.8.0.
BNC Consortium. 2007. British National Corpus (BNC). Online: http://www.natcorpus.ox.ac.uk
Bowen, Matt. 2019. Stigma: A linguistic analysis of personality disorder in the UK popular press, 2008-2017. *Journal of Psychiatry Mental Health Nursing,* 26(7-8), 244-253. Wiley Publishing.
Brezina, Vaclav, Tony McEnery, and Stephen Wattam. 2015. Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics* 20:2*,* 139-173. John Benjamins Publishing Company.
Brezina, Vaclav, Robbie Love, and Karin Aijmer. 2018. *Corpus Approaches to Contemporary British Speech: Sociolinguistic Studies of the Spoken BNC2014.*
Brezina, Vaclav. 2018. *Statistics in Corpus Linguistics: A Practical Guide.* Cambridge University Press.
Brittanica. 2021. World Health Organization: UN Public Health Agency. Britannica Editors. Britannica Online.
British Academy. 2020. Vaccine hesitancy threatens to undermine pandemic response. Online: https://www.thebritishacademy.ac.uk/news/vaccine-hesitancy-threatens-undermine-pandemic-response/?utm_source=twitter&utm_medium=social&utm_campaign=news%20%7C%20announcement%20%7C%20%20%7C%20Press&utm_content=Press&utm_term=20201110

Broniatowski, David, Michael Paul, and Mark Dredze. 2013. National and Local Influenza Surveillance through Twitter: An Analysis of the 2012-2013 Influenza Epidemic. *PLoS ONE*.

Brookes, Gavin, Kevin Harvey, and Louise Mullany. 2018. From corpus to clinic: Health communication research and the impact agenda. *Applying Linguistics: Language and the Impact Agenda, ed.* Dan McIntyre and Hazel Price. New York: Routledge.

Brookes, Gavin, and Tony McEnery. 2019. The utility of topic modelling for discourse studies: A critical evaluation. *Discourse Studies.*

Brown, Brian, Paul Crawford, and Ronald Carter. 2006. *Evidence-Based Health Communication.* McGrew-Hill Education. In Brookes et al., 109.

Burkette, Allison, and William Kretzschmar. 2018. *Exploring Linguistic Science: Language Use, Complexity and Interaction.* Cambridge University Press.

Burrows, John. 2004. Textual Analysis. *A New Companion to Digital Humanities,* ed. by Susan Schriebman, Ray Siemens, and John Unsworth. John Wiley & Sons Publishing.

Burton, Robert. 1621. *Anatomy of Melancholy*. In Wear, 127.

Bylund, Carma L. and Christopher J. Koenig. 2015. Approaches to Studying Provider-Patient Communication. *Health Communication: Theory, Method, and Application,* ed. by Nancy Grant Harrington. 116-146. Routledge.

Bynum, William, Browne, E. Janet, & Porter, Roy. 2014. *Dictionary of the History of Science*. Princeton: Princeton University Press.

Casaa.org. 2012-2019. Historical Timeline of Electronic Cigarettes. Date Accessed: 5 Oct 2019.

Chen, Alvin. 2020. Chapter 6: Keyword Analysis. Online: https://alvinntnu.github.io/NTNU_ENC2036_LECTURES/keyword-analysis.html#ref-gries2018

Cheng, Winnie. 2009. Describing the extended meaning of lexical cohesion in a corpus of SARS spoken discourse. *Lexical Cohesion and Corpus Linguistics,* ed. by John Flowerdew and Michaela Mahlberg. John Benjamins Publishing.

Clarke, Isobella and Jack Grieve. 2017. Dimensions of Abusive Language on Twitter. in *Proceedings of the First Workshop on Abusive Language Online. Association for Computational Linguistics,* 1-10.

Clarke, Isobella, and Jack Grieve. 2019. Stylistic variation on the Donald Trump Twitter account: a Linguistic Analysis of tweets posted between 2009 and 2018. *PLoS ONE* 14(9).

Clear, J. 1992. Corpus sampling. *New directions in English language corpora,* ed. G. Leitner, 21-31. Berlin: Mouton de Gruyter. In Sinclair, Chapter 1: Corpus and Text – Basic Principles

Collins, Luke and Brigitte Nerlich. 2016. Uncertainty discourses in the context of climate change: A corpus-assisted analysis of UK national newspaper articles. *Communications,* 291-313. De Gruyter Mouton.

Coleman, William. 1982. *Death is a Social Disease: Public Health and Political Economy in Early Industrial France.* The University of Wisconsin Press.

Conrad, Lawrence, Michael Neve, Vivian Nutton, Roy Porter, and Andrew Wear. 1995. *The Western Medical Tradition: 800 BC to AD 1800.* Cambridge, UK: Cambridge University Press.

Crespo, Begoña, and Isabel Moskowich. 2016. At Close Range: Prefaces and Other Text Types in the *Coruña Corpus of English Scientific Writing.* Revista de Lenguas para Fines Especificos 22(10, 213-237. Universidad de Las Palmas de Gran Canaria.

Cueto, Marcos, Theodore M. Brown, and Elizabeth Fee. 2019. *The World Health Organization: A History.* Cambridge University Press.

Culpepper, Jonathan. 2009. Keyness: Words, parts-of-speech and semantic categories in the character talk of Shakespeare's Romeo and Juliet. *International Journal of Corpus Linguistics.* 14.1. John Benjamins Publishing.

Curth, L. Hill. 2003. Lessons from the past: preventive medicine in early modern England. *Journal of Medical Ethics: Medical Humanities,* 16-21.
Damerau, Fred J. 1993. Generating and Evaluating Domain-Oriented Multi-Word Terms from Texts. *Information Processing & Management* 29 (4), 433–47.
Davies, Mark. 2008. *The Corpus of Contemporary American English.* Online: *https://www.english-corpora.org/coca/.*
Davies, Mark. 2010. *The Corpus of Historical American English (COHA)*. Online: https://www.english-corpora.org/coha/.
Delamothe, Tony. 2008. Founding principles: NHS at 60. *British Medical Journal,* 1216-1218.
De Renzi, Silvia. 2004. Old and New Models of the Body. *The Healing Arts: Health, Disease and Society in Europe: 1500-1800,* ed. by Peter Elmer, 166-192. Manchester, UK: Manchester University Press.
De Renzi, Silvia. 2004. The sick and their healers. *The Healing Arts: Health, Disease and Society in Europe: 1500-1800,* ed. by Peter Elmer, 27-25. Manchester, UK: Manchester University Press.
Diamond, Larry. 2012. Introduction: *Liberation Technology, Social Media and the Struggle for Democracy,* ed. Larry Diamond and Marc Plattner. The Johns Hopkins University Press.
Donovan, Arthur. 1988. Lavoisier and the Origins of Modern Chemistry. *History of Science Society* (4): 214- 231. The University of Chicago Press.
Douthat, Ross. 2020. The End of the New World Order. *New York Times.* 24 May 2020.
Duggan, M. Ellison, N. Lampe, C, Lenhart, A, and Madden, M. 2015. *Social Media update 2014.* Pew Research Center.
Dunder, I., Horvat, M., & Lugovic, S. 2016. Word occurrences and emotions in social media: Case study on a Twitter corpus. *39th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO).*
Dunning, Ted. 1993. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics 19 (1)*: 61–74. Online: *https://www.aclweb.org/anthology/J93-1003.*
Eisenstein, Jacob, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2014. Diffusion of Lexical Change in Social Media. *PLOS ONE* 9(11).
Etheridge, Elizabeth. 1992. *Sentinel for Health: A History of the Centers for Disease Control.* Los Angeles: University of California Press.
Etter, Jean-Francois and Chris Bullen. 2011. Electronic cigarette: users profile, utilization, satisfaction and perceived efficacy. *Addiction.*
Evert, Stefan and Andrew Hardie. 2011. Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. *Proceedings of the Corpus Linguistics 2011 conference,* University of Birmingham, UK.
Evert, Stefan, and the CWB Development Team. 2021. The IMS Open Corpus Workbench (CWB) Corpus Encoding and Management Manual. Online: http://cwb.sourceforge.net/
Fairclough, Norman. 2003. *Analysing Discourse: Textual Analysis for Social Research.* Routledge Publishing.
Falk, Seb. 2020. *The Light Ages: The Surprising Story of Medieval Science.* W. W. Norton and Company.
Feinerer, I. and K. Hornik. 2019. tm: Text Mining Package. R package version 0.7-7.
Feinerer, Ingo. 2020. Introduction to the tm Package: Text Mining in R. https://cran.r-project.org/web/packages/tm/vignettes/tm.pdf
Firth, J.R. 1935. The Technique of Semantics. *Transactions of the Philological Society*, 36-72.
Firth, J. R. 1957. *Papers in Linguistics 1934-1951.* London: Oxford University Press.
Fischer, Stefan, Jörg Knappen, Katrin Menzel, Elke Teich. 2020. The Royal Society Corpus 6.0: Providing 300+ Years of Scientific Writing for Humanistic Study. *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020),* 794-801.

Fissell, Mary E. 1991. *Patients, Power, and the poor in Eighteenth Century Bristol.* Cambridge, UK: Cambridge University Press.
Fox, Nick J. 1993. *Postmodernism, Sociology and Health.* Buckingham: Open University Press. In Brookes et al. 2018, 99.
Francis, Gill, and Anneliese Kramer-Dahl. 2004. Grammar in the Construction of Medical Case Histories. *Applying English Grammar: Functional and Corpus Approaches,* ed. by Caroline Coffin, Ann Hewings, and Kieran O'Halloran, 172-190. Hodder Arnold. In Atkins & Harvey, 606.
Franklin, Bob. 2016. *The Future of Journalism: In an Age of Digital Media and Economic Uncertainty.* Routledge Publishing.
Friedland, Gerald. 2009. Discovery of the function of the heart and circulation of blood. *Cardiovascular Journal of Africa,* 160. *US National Institutes of Health.*
Gabrielatos, Costas and Paul Baker. 2008. Fleeing, Sneaking, Flooding: A Corpus Analysis of Discursive Constructions of Refugees and Asylum Seekers in the UK Press, 1996-2005. *Journal of English Linguistics* 36(1). SAGE Publishing.
Gabrielatos, Costas & Anna Marchi. 2012. Keyness: Appropriate metrics and practical issues: CADS 2012.
Gabrielatos, Costas. 2018. Chapter 12: Keyness Analysis: nature, metrics and techniques. *Corpus Approaches To Discourse: A critical review,* ed. by Charlotte Taylor and Anna Marchi. Routledge Publishing.
Garner, James, Scott Crossley and Kristopher Kyle. 2018. Beginning and intermediate L2 writer's use of N-grams: an association measures study. *International Review of Applied Linguistics in Language Teaching.*
Gentilcore, David. 1998. *Healers and Healing in Early Modern Italy.* Manchester, UK: Manchester University Press.
Gläser, Rosemarie. 1995. *Linguistic Features and Genre Profiles of Scientific English.* Berlin: Peter Lang.
Goddard, Cliff and Anna Wierzbicka. 2014. *Words and Meaning: Lexical Semantics Across Domains, Languages, and Cultures.* Oxford University Press.
Goel, Rahul, Sandeep Soni, Nama Goyal, John Paparrizos, Hanna Wallach, Fernando Diaz, and Jacob Eisenstein. 2016. The Social Dynamics of Language Change in Online Networks. *Proceedings of the International Conference on Social Informatics (SocInfo16).*
Görlach, Manfred. 1999. *English in Nineteenth-Century England: An introduction.* Cambridge University Press.
Gotti, Maurizio. 2011. The development of specialized discourse in the Philosophical Transactions. *Medical Writing in Early Modern English,* ed. by Irma Taavitsainen and Päivi Pahta. Cambridge University Press.
Götze, Heinz. 1995. The English Language in Scientific Publishing. *Publishing Research Quarterly* 13. 52-71.
Gray, Bethany, Douglas Biber, and Turo Hiltunen. 2011. The expression of stance in early (1665-1712) publications of the *Philosophical Transactions* and other contemporary medical prose: innovations in a pioneering discourse. *Medical Writing in Early Modern English,* ed. by Irma Taavitsainen and Päivi Pahta. Cambridge University Press.
Greaves, Chris and Martin Warren. 2010. What can a corpus tell us about multi-word units? In *Routledge Handbook of Corpus Linguistics,* ed. by Anne O'Keeffe and Michael McCarthy. London: Routledge Publishing.
Gries, Stefan. 2009. What is Corpus Linguistics? *Language and Linguistics Compass* 3. 1-17.
Gries, Stefan Th. 2013. 50-something years of work on collocations: What is or should be next… *International Journal of Corpus Linguistics* 18:1, 137-165. John Benjamins Publishing Company.

Gries, Stefan. 2017. *Quantitative Corpus Linguistics with R: A Practical Introduction: 2nd Edition.* Routledge.
Grieve, Jack, Andrea Nin, and Diansheng Guo. 2018. Mapping Lexical Innovation on American Social Media. *Journal of English Linguistics* 46(4), 293–319. SAGE Publishing.
Grieve, Jack, Chris Montgomery, Andrea Nini, Akira Murakami, and Diansheng Guo. 2019. Mapping Lexical Dialect Variation in British English Using Twitter. *Frontiers of Artificial Intelligence* 2:11.
Gross, Alan, Joseph Harmon, and Michael Reidy. 2002. *Communicating Science: The Scientific Article from the 17th Century to the Present.* Oxford, UK: Oxford University Press.
Gui, X., Wang, Y., Kou, Y., Reynolds, T. L., Chen, Y., Mei, Q., & Zheng, K. 2018. Understanding the Patterns of Health Information Dissemination on Social Media during the Zika Outbreak. *AMIA Annual Symposium proceedings,* 820–829. AMIA Symposium.
Gupta, Kat. 2015. *Representation of the British Suffrage Movement: Research in Corpus and Discourse,* ed. by Michaela Mahlberg & Wolfgang Teubert. Bloomsbury Publishing.
Haider, Muhiuddin and Everett M. Rogers. 2005. Introduction: Public Health Communication: Utility, Values, and Challenges. *Global Public Health Communication: Challenges, Perspectives, and Strategies,* ed. by Haider Muhiuddin. Jones and Bartlett Publishers.
Hannaford, Ewan. 2017. The Press and the Public Attitudes on Mental Health: A Corpus Linguistic Analysis of UK Newspaper Coverage of Mental Illness (1994-2014), Compared with the UK National Attitudes to Mental Illness Survey. MPhil Thesis: University of Glasgow.
Hannaford, Ewan. 2019. Questioning the mental/physical health divide: Investigating illness discourses in the press using corpus linguistics. American Association of Applied Linguistics.
Hannaford, Ewan. 2021. *Interpreting illness and the mental/physical health divide: A corpus linguistic investigation of illness representations in the UK and US press (1995-2017).* University of Glasgow. PhD Thesis.
Hardie, Andrew. 2012. CQP Web: combining power, flexibility, and usability in a corpus analysis tool. *International Journal of Corpus Linguistics.* 380-409. John Benjamins Publishing.
Harrington, Nancy Grant. 2015. *Health Communication: Theory, Method, and Application.* Routledge Publishing.
Hazen, Kirk. 2014. *An Introduction to Language.* Wiley Publishing.
Hiltunen, Turo, and Jukka Tyrkkö. 2011. Verbs of knowing: discursive practices in early modern vernacular medicine. *Medical Writing in Early Modern English,* ed. by Irma Taavitsainen and Päivi Pahta. Cambridge University Press.
Hunston, Susan. 2010. How Can a Corpus Be Used to Explore Patterns? *Routledge Handbook of Corpus Linguistics,* ed. by Anne O'Keeffe and Michael McCarthy. London: Routledge Publishing.
HKSAR. 2014. The Hong Kong Special Administrative Region. Ministry of Foreign Affairs for the People's Republic of China. https://www.fmprc.gov.cn/mfa_eng/ljzg_665465/zgjk_665467/3572_665469/t17814.shtml
Honigsbaum, Mark. 2019. *The Pandemic Century: One Hundred Years of Panic, Hysteria, and Hubris.* London: Norton and Company.
Ho, Evelyn Y. 2015. Socio-cultural Factors in Health Communication. *Health Communication: Theory, Method, and Application,* ed. by Nany Harrington. Routledge Publishing.
Hornik, Robert. 2002. Public Health Communication: Evidence for Behavior Change. *Taylor & Francis.*
Hunston, Susan. 2012. *Corpora in Applied Linguistics.* Cambridge Applied Linguistics. Cambridge University Press.

Hunt, Daniel and Kevin Harvey. 2015. Health Communication and Corpus Linguistics: Using Corpus Tools to Analyse Eating Disorder Discourse Online. *Corpora and Discourse Studies*. Palgrave Advances in Language and Linguistics.
Hunt, Daniel and Gavin Brookes. 2020. *Corpus, Discourse and Mental Health*. Bloomsbury Publishing.
Jacobs, Lawrence R. 1993. *The Health of the Nations: Public Opinion and the Making of American and British Health Policy.* Cornell University Press: Ithaca and London.
Jaspal, Rusi, and Brigitte Nerlich. 2017. Polarised press reporting about HIV prevention: Social representations of pre-exposure prophylaxis in the UK press. *Health.* Vol 21:50. 478-497. SAGE Publishing.
Jenner, Edward. 2004. Environment, Health, and Population. *The Healing Arts: Health, Disease and Society in Europe: 1500-1800,* ed. by Peter Elmer, 284-314. Manchester, UK: Manchester University Press.
Jockers, Matthew. 2013. *Macroanalysis: Digital Methods and Literary History.* University of Illinois Press.
Jockers, Matthew, and Julia Flanders. 2013. A Matter of Scale. Keynote Lecture from the Boston Area Days of Digital Humanities Conference. Northeastern University, Boston, MA. March 18, 2013.
Jockers, Matthew and Ted Underwood. 2015. Textmining the humanities. *A New Companion to Digital Humanities,* ed. by Susan Schriebman, Ray Siemens, and John Unsworth. John Wiley & Sons Publishing.
Jockers, Matthew. 2016. The Ancient World in 19th century fiction; or, Correlating Theme, Geography, and Sentiment in the 19th Century Literary Imagination. *Digital Humanities Quarterly* 10.2. The Association for Computers and the Humanities.
Juyal, Deepak, Shekhar Pal, Shweta Thaledi, Shalabh Jauhari, and Sunil Kumar. 2020. Role of social media amidst coronavirus disease 2019 crisis: Expectations versus reality. *World Journal of Nuclear Medicine.*
Kabat, Geoffrey. 2017. *Getting Risk Right: Understanding the Science of Elusive Health Risks.* Columbia University Press.
Kamadjeu, Raoul. 2019. English: the *lingua franca* of scientific research. *The Lancet: Global Health* 7:9.
Kaplan, Robert. 2001. English – the Accidental Language of Science? *The Dominance of English as a Language of Science: Effects on Other Languages and Language Communities,* ed. by Ulrich Ammon. Mouton de Gruyter.
Kermes, Hannah, Stefania Degaetano-Ortleib, Ashraf Khamis, Jörg Knappen, and Elke Teich. 2015. The Royal Society Corpus: From Uncharted Data to Corpus.
Kipple, Kenneth F. 2006. The History of Disease. *The Cambridge History of Medicine,* ed. by Roy Porter. Cambridge, UK: Cambridge University Press.
Kretzschmar, William, C. Darwin, C. Brown, D. Rubin, D. Biber. 2004. Looking for the Smoking Gun: Principled Sampling in Creating the Tobacco Industry Documents Corpus. *Journal of English Linguistics* 32:1. SAGE Publishing.
Kretzschmar, William. 2009. The Linguistics of Speech. Cambridge University Press.
Kretzschmar, William. 2015. *Language and Complex Systems.* Cambridge University Press.
Kretzschmar, William. 2018. *The Emergence and Development of English.* Cambridge University Press.
Kretzschmar, William. *Forthcoming.* Complex Systems for Corpus Linguists. *ICAME Journal.*
Kuiper, Katherine Ireland. 2020. The E-cigarette Dilemma. *ICAME* 41.
Kuiper, Katherine Ireland. *Forthcoming.* Keywords and diachronic analysis in the Royal Society Corpus.

Kusukawa, Sachiko. 2004. Medicine in Western Europe in 1500. *The Healing Arts: Health, Disease and Society in Europe: 1500-1800,* ed. by Peter Elmer, 1-24. Manchester, UK: Manchester University Press.
Lamsal R. 2020. Coronavirus (COVID-19) Tweets Dataset. IEEE Dataport.
Lane, Nick. 2015. The unseen world: reflections on Leeuwenhoek (1677) 'Concerning little animals'. *Philosophical Transactions of the Royal Society of London B Biological Sciences:* 370.
Larson, Heidi. 2020. *Stuck: How Vaccine Rumors Start-and Why They Don't Go Away.* Oxford University Press.
Leech, Geoffrey, and Roger Fallon. 1992. Computer Corpora–What Do They Tell Us About Culture. *ICAME Journal* 16.
Leong, Alvin. 2020. The passive voice in scientific writing through the ages: A diachronic study. *Text and Talk,* 467-489. De Gruyter Mouton.
Livingstone, David N. 2003. *Putting Science in its Place: Geographies of Scientific Knowledge.* Chicago and London: The University of Chicago Press.
Maibach, E. and D. R. Holtgrave. 1995. Advances in Public Health Communication. *Annual Review of Public Health* 16, 219-38.
Mair, Christian. 2005. *Twentieth Century English.* Cambridge University Press.
Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David Marcus, Mitchell P., Beatrice Santorini, Mary Ann Marcinkiewicz. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *Association for Computational Linguistics.*
Matson, Cathy. 2019. Foreword. *Revolutionary Networks: the Business and Politics of Printing the News, 1763-1789.* Baltimore: Johns Hopkins University Press.
Martel, Angéline. 2001. When Does Knowledge Have a National Language? Language Policy-Making for Science and Technology. *The Dominance of English as a Language of Science: Effects on Other Languages and Language Communities,* ed. Ulrich Ammon. Mouton de Gruyter.
Mautner, Gerlinde. 2008. Analyzing Newspapers, Magazines and Other Print Media. *Qualitative Discourse analysis in the Social Sciences,* ed. Ruth Wodak and Michal Krzyżanowski. Macmillan.
Merrick, Thomas W. 2005. Forward. *Global Public Health Communication: Challenges, Perspectives, and Strategies,* ed. Haider Muhiuddin. Jones and Bartlett Publishers.
McCallum, Andrew Kachites. 2002. MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu
McChonchie, Ron and Anne Curzan. 2011. Defining in early Modern English medical texts. In *Medical Writing in Early Modern English,* ed. Irma Taavitsainen and Päivi Pahta. Cambridge University Press.
McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 55-60.
McEnery, Tony and Helen Baker. 2017. *Corpus Linguistics and 17th Century Prostitution*. Bloomsbury Publishing.
McEnery, Tony, Vaclav Brezina, and Helen Baker. 2019. Usage Fluctuation Analysis: A new way of analysing shifts in historical discourse. *International Journal of Corpus Linguistics* 24(4). John Benjamins Publishing.
McEvoy, John G. 1992. The Chemical Revolution in Context. *The Eighteenth Century,* 198-216. University of Pennsylvania Press.
McIntosh, Christopher. 2011. The Phlogiston Theory: A Late Reliv of Pre-Enlightenment Science. *Handbook of Religion and the Authority of Science,* ed. by Jim Lewis and Olav Hammer Brill Handbooks on Contemporary Religion (3).

McKinney, Wes. 2010. Data structures for statistical computing in python. *Proceedings of the 9th Python in Science Conference,* ed. by Stefan van der Walt and Jarrod Millman.
Millar, P, and BS Budgell. 2008. The language of public health-a corpus-based analysis. *Journal of Public Health* 16(5). Springer-Verlag Publishing.
Mold, Alex, and Virginia Berridge. 2018. Using Digitized Medical Journals in a Cross European Project on Addiction History. *Media History* 25, 85-99. Taylor and Francis.
Monaco, Leida. 2016. Was late Modern English scientific writing impersonal? Comparing Philosophy and Life Sciences texts from the Coruña Corpus. *International Journal of Corpus Linguistics,* 499-526. John Benjamins Publishing.
Morley, J. & Taylor, C. 2012. Us and them: How immigrants are constructed in British and Italian newspapers. *European Identity: What the Media Say,* ed. by P. Bayley & G. Williams, 190–223. Oxford: Oxford University Press.
Moskowich, Isabel, Luis Puente-Castelo, Begoña Crespo-Garcia, Gonzalo Camiña-Rioboó. 2021. The Coruña Corpus of English Scientific Writing: Challenge and Reward.
Mowery, Danielle, Hilary Smith, Tyler Cheney, Greg Stoddard, Glen Coppersmith, Craig Bryan, and Mike Conway. 2017. Understanding Depressive Symptoms and Psychosocial Stressors on Twitter: A Corpus-Based Study. *Journal of Medical Internet Research.* JMIR Publications.
Newton, Jennifer. 2020. Britons reveal the top 50 things that make them proud to be British. *Dail Mail.* Online: *https://www.dailymail.co.uk/travel/travel_news/article-8180151/Britons-reveal-50-things-make-proud-British-NHS-David-Attenborough.html*
NHS. 2021. NHS COVID-19 app support. *NHS Test and Trace.* Online: *https://covid19.nhs.uk/*
O'Donnell, Matthew Brook, Ute Römer, and Nick Ellis. 2013. The development of formulaic sequences in first and second language writing: Investigating effects of frequency association and native norm. *International Journal of Corpus Linguistics* 18:1, 83-108. John Benjamins Publishing.
Oliver del Olmo, Sonia. 2014. Hedging and attitude markers in Spanish and English scientific medical writing. *Communicating Certainty and Uncertainty in Medical, Supportive and Scientific Contexts,* ed. by Andrzej Suczkowski, Ramona Bongelli, Ilaria Riccioni, Carla Canestrari. John Benjamins Publishing.
pandas developers. 2021. pandas. version 1.2.3. Online: *https://pandas.pydata.org/docs/*
Ovenden, Richard. 2020. *Burning the Books: A History of The Deliberate Destruction of Knowledge.* Belnap Press.
Pano, Toni, and Rasha Kashef. 2020. A Corpus of BTC Tweets in the Era of COVID-19. *IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS).* IEEE International.
Patra, B.G., Nilabjya Ghosh, Diparikar Das, and Sivaji Bandyopadhijay. 2015. Identifying Temporal Information and Tracking Sentiment Analysis in Cancer Patients Interviews. Computational Linguistics and Intelligent Text Processing: 18th CICLing Proceedings Part 2, 180-188. Springer Publishing.
Park, Hyojung, Bryan H Reber, and Myoung-Gi Chon. 2016. Tweeting as Health Communication: Health Organizations' Use of Twitter for Health Promotion and Public Engagement. *Journal of Health Communication.* National Library of Medicine.
Parrott, Roxanne, and Kreuter, Matthew W. 2011. Multidisciplinary, Interdisciplinary, and Transdisciplinary Approaches to Health Communication: Where Do We Draw the Lines? *The Routledge Handbook of Health Communication,* ed. by Teresa L. Thompson, Roxanne Parrott, Jon F. Nussbaum. Routledge Publishing.
Partington, Alan. 2012. The changing discourses on antisemitism in the UK press from 1993 to 2009: A modern-diachronic corpus-assisted discourse study. *Journal of Language and Politics* 11:1: 51-76. John Benjamins Publishing.

Paterson, Laura and Ian Gregory. 2019. *Representations of Poverty and Place: Using Geographical Text Analysis to Understand Discourse.* Springer Publishing.
Patra, B.G., Nilabjya Ghosh, Diparikar Das, and Sivaji Bandyopadhijay. 2015. Identifying Temporal Information and Tracking Sentiment Analysis in Cancer Patients Interviews. *Computational Linguistics and Intelligent Text Processing: 18th CICLing Proceedings Part 2*, 180-188. Springer Publishing.
Pender, Stephen. 2010. Subventing Disease. *Rhetorics of Bodily Disease and Health in Medieval and Early Modern England,* ed. Jennifer Vaught, 193-218. Ashgate Publishing.
Potts, Amanda, and Elena Semino. 2017. Healthcare professionals' online use of violence metaphors for care at the end of life in the US: a corpus-based comparison with the UK. *Corpora.* Vol 12.1, 55-84. Edinburgh: Edinburgh University Press.
Porter, Roy. 1989. *Health for Sale: Quackery in England 1660-1850.* Manchester and New York: Manchester University Press.
Porter, Roy. 2000. *The Creation of the Modern World: The Untold Story of the British Enlightenment.* New York: W.W. Norton & Company.
Porter, Roy. 2001. *The Enlightenment.* Palgrave Publishing.
Porter, Roy. 2006a. *The Cambridge History of Medicine.* Cambridge, UK: Cambridge University Press.
Porter, Roy. 2006b. Introduction. *The Cambridge History of Medicine,* ed. Roy Porter. Cambridge, UK: Cambridge University Press.
Porter, Roy. 2006c. What is Disease? *The Cambridge History of Medicine,* ed. Roy Porter. Cambridge, UK: Cambridge University Press.
Porter, Roy. 2006d. Medical Science. *The Cambridge History of Medicine,* ed. Roy Porter. Cambridge, UK: Cambridge University Press.
Porter, Roy. 2006e. Hospitals and Surgery. *The Cambridge History of Medicine,* ed. Roy Porter. Cambridge, UK: Cambridge University Press.
Porter, Roy. 2006f. Mental Illness. *The Cambridge History of Medicine,* ed. Roy Porter. Cambridge, UK: Cambridge University Press.
Python. Version 3.7.4. July 8 2019.
Real, Kevin, and Marjorie M. Buckner. 2015. Interprofessional Communication: Health Care Teams and Medical Interpreters. *Health Communication: Theory, Method, and Application,* ed. by Nancy Grant Harrington. 147-178. Routledge.
R Core Team 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria.
Regidor, Enrique, Cruz Pascual, Luis de la Fuente, Salvador de Matea, Juan Gutierrez-Fisac, José Sanchez-Paya, Elena Ronda. 2007. The Role of the Public Health Official in Communicating Public Health Information. *American Journal of Public Health.*
Ringrow, Helen. 2016. *The Language of Cosmetics Advertising.* London: Palgrave Macmillan.
Reiser, Stanley Joel. 1978. *Medicine and the reign of technology.* Cambridge: Cambridge University Press.
Römer, Ute. 2010. Establishing the phraseological profile of a text type: The construction of meaning in academic book reviews. *English Text Construction.* John Benjamins Publishing.
Rosenberg, Charles. 2007. Foreword. *The Making of a Tropical Disease: a Short History of Malaria.* Johns Hopkins University Press.
Royal College of Physicians. 2016. *Nicotine without smoke: Tobacco harm reduction.* London: RCP.
The Royal Society. 2020. History of the Royal Society. Online: https://royalsociety.org/about-us/history/
Rüdiger, Sophia, and Daria Dayter. 2020. Corpus Approaches to Social Media. *Studies in Corpus Linguistics.* John Benjamins Publishing.

Saunders, Patrick, Andrew Kibble, and Amanda Burls. 2006. Investigating Alleged Clusters. In *Oxford Handbook of Public Health Practice: 2nd Edition,* ed. by David Pencheon, Charles Guest, David Melzer, and J.A. Muir Gray. Oxford, UK: Oxford University Press.

Schuchat, Anne, & CDC COVID-19 Response Team. 2020. Public Health Response to the Initial Spread of the Pandemic COVID-19 in the United States. *Morbidity and Mortality Weekly Report: MMRW.* 551-556.

Scott, Mike. 2010. What can corpus software do? *Routledge Handbook of Corpus Linguistics,* ed. by Anne O'Keeffe and Michael McCarthy. London: Routledge Publishing.

Scott, Mike and Christopher Tribble. 2006. *Textual Patterns: Key words and corpus analysis in language education.* John Benjamins Publishing.

Schmid, Helmut. 1995. Improvements in Part-of-Speech Tagging with an Application to German. *Proceedings of the ACL SIGDAT-Workshop.* Dublin, Ireland.

Semino, Elena, Zsófia Demnjén, Jane Demmen, Veronika Koller, Sheila Payne, Andrew Hardie, and Paul Rayson. 2015. The online use of Violence and Journey metaphors by patients with cancer, as compared with health professionals: a mixed methods study. *BMJ Supportive and Palliative Care*.

Semino, Elena, Sakrzewska, JM, Williams, A. 2017. Images and the dynamics of pain consultations. *The Lancet.*

Semino, Elena, Zsófia Demnjén, Jane Demmen. 2018. An integrated approach to metaphor and framing in cognition, discourse and practice, with an application to metaphors for cancer. *Applied Linguistics* 39:5. 625-645.

Semino, Elena and Veronic Koller. 2020. Talking about lockdown language. Lancaster University Public Lecture.

Shyrock, Richard Harrison. 1947. *The Development of Modern Medicine: An Interpretation of the Social and Scientific Factors Involved.* New York: Alfred Knopf Publishing.

Siegel, Michael, Kerry Tanwar, and Kathleen Wood. 2011. Electronic Cigarettes as a Smoking-Cessation Tool: Results from an Online Survey. *American Journal of Preventive Medicine.* Elsevier Inc.

Silge, Julia, and David Robinson. 2016. Tidytext: Text Mining and Analytics Using Tidy Data Principles in R. *The Open Journal* 1:3.

Silge, Julia, and David Robinson. 2021. Text Mining with R: A Tidy Approach. Online: https://www.tidytextmining.com/

Sinclair, John. 2004. *Trust the Text: Language corpus and discourse.* Routledge Publishing: London, UK.

Sinclair, John. 2004. Chapter 1: Corpus and Text – Basic Principles. *Developing Linguistic Corpora: A Guide to Good Practice,* ed. by Martin Wynne. AHDS: Literature, Language and Linguistics.

Skelton, J. and Hobbs, F.D.R. 1999. Concordancing: the use of language-based research in medical communication. *The Lancet*, 353: 108-111.

Skelton, J.R., Wearn, A. M., and F.D.R. Hobbs. 2002. 'I' and ' we': a concordancing analysis of doctors and patients use first person pronouns in primary care consultations. *Family* Practice (19)5: 484-488.

Smith, Matthew. 2018. The NHS is the British institution that Brits are second-most proud of – after the fire brigae. YouGov UK PLC. Online: https://yougov.co.uk/topics/politics/articles-reports/2018/07/04/nhs-british-institution-brits-are-second-most-prou

Stubbs, M. 1995. Collocations and cultural connotations of common words. *Linguistics and Education.* 370-390. Elsevier.

Stubbs, Michael. 1995. Collocations and Semantic Profiles: on the Cause of the Trouble with Quantitative Studies. Reprinted in 2007 in Teubert W. and R. Krishnamurthy (eds) *Corpus Linguistics: Critical Concepts in Linguistics.* London and New York: Routledge.

Stubbs, Michael. 1996. *Text and Corpus Analysis.* Blackwell Publishing.
Stubbs, Michael. 2001. *Words and Phrases: Corpus Studies of Lexical Semantics.* Blackwell Publishing.
Taavitsainen, Irma, and Pahta. 1998. Vernacularization of medical writing in English: A corpus-based study of scholasticism. In *Early Science and Medicine. Brill.* 157-85.
Taavitsainen, Irma et al. 2000. *Corpus of Early English Medical Writing 1375–1800.*
Taavitsainen, I, and Pahta P. 2009. The pragmatics of knowledge and meaning: Corpus linguistic approaches to changing thought styles in early modern medical discourse. In *Corpora, pragmatics, and discourse,* Ed. Jucker, A., Schreier, D., and Hundt, M. *Rodopi.*
Taavitsainen, Irma and Päivi Pahta. 2011. An interdisciplinary approach to medical writing in Early Modern English. *Medical Writing in Early Modern English,* ed. by Irma Taavitsainen and Päivi Pahta. Cambridge University Press.
Taavitsainen, Irma, Peter Murray Jones, Päivi Pahta, Turo Hiltunen, Ville Marttila, Maura Ratia, Carla Suhr, and Jukka Tyrkkö. 2011. Medical texts in 1500-1700 and the corpus of *Early Modern English Medical Texts. Medical Writing in Early Modern English,* ed. by Irma Taavitsainen and Päivi Pahta. Cambridge University Press.
Taavitsainen, Irma. 2011. Dissemination and appropriation of medical knowledge: humoral theory in Early Modern English medical writing and lay texts. *Medical Writing in Early Modern English,* ed. by Irma Taavitsainen and Päivi Pahta. Cambridge University Press.
Taavitsainen, Irma, Turo Hiltunen, Anu Lehto, Ville Marttila, Päivi Pahta, Maura Ratia, Carla Suhr, and Jukka Tyrkko. 2014. Late Modern English Texts 1700-1800: A corpus for analysing eighteenth-century medical English. *ICAME Journal.*
Taavitsainen, Irma. 2016. Genre dynamics in the history of English. *Cambridge Handbook of English Historical Linguistics,* ed. by Merja Kytö and Päivi Pahta. Cambridge University Press.
Taavitsainen, Irma. 2017. Meaning-making practices in the history of medical English: A sociopragmatic approach. *Journal of Historical Pragmatics.*
Taavitsainen, Irma, and Gerold Schneider. 2019. Scholastic Argumentation in Early English Medical Writing and Its Afterlife: New Corpus Evidence. *From Data to Evidence in English Language Research.* 191-221.
Tang, Chris, and Gabriella Rundblad. 2017. When Safe Means 'Dangerous': A Corpus Investigation of Risk Communication in the Media. *Applied Linguistics.* Oxford University Press.
Tassava, Christopher. 2008. The American Economy during World War Two. Economic History Association. Online: EH.net.
Taylor, Charlotte. 2014. Investigating the representation of migrants in the UK and Italian press: A cross-linguistic corpus-assisted discourse analysis. *International Journal of Corpus Linguistics.* 368-400. John Benjamins Publishing.
Taylor, Derrick Bryson. 2021. A Timeline of the Coronavirus Pandemic. *New York Times.* Online: *https://www.nytimes.com/article/coronavirus-timeline.html*
Taavitsainen, I, and Pahta P. 1998. Vernacularization of medical writing in English: A corpus-based study of scholasticism. *Early Science and Medicine,* 157-185. *Brill.*
Teubert, Wolfgang. 2019. Corpus linguistics: Widening the remit. *Corpus Linguistics, Context and Culture,* ed. by Viola Wiegand and Michaela Mahlberg. De Gruyter Publishing.
Thomas, J. and Wilson, A. 1996. Methodologies for studying a corpus of doctor-patient interaction. *Using Corpora for Language Research*, ed. by J. Thomas and M. Short. M.London: Longman.
Tinniswood, Adrian. 2019. *The Royal Society.* London: Apollo Publishing.
Toutanova, Kristina, Dan Klein, Christopher Manning, and Yoram Singer. 2003. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. *Proceedings of HLT-NAACL 2003*, 252-259.

Tufekci, Zeynep. 2017. *Twitter and Tear Gas: The Power and Fragility of Networked Protest.* Yale University Press.
Underwood, Ted. 2019. Algorithmic Modeling. *The Shape of Data in Digital Humanities: Modeling Texts and Text-based Resources,* ed. by Julia Flanders and Fotis Jannidis. Routledge.
Veil, Shari R. and Timothy L. Sellnow. 2015. Risk and Crisis Communication. *Health Communication: Theory, Method, and Application,* ed. by Nany Harrington. Routledge.
Walley, SC, KM Wilson, JP Winickoff, and J Groner. 2019. A Public Health Crisis: Electronic Cigarettes, Vape, and JUUL. *Pediatrics.*
Wang, Ying. 2017. Lexical bundles in news discourse 1784-1983. In *Diachronic Developments in English News Discourse.* John Benjamins Publishing.
Whitelaw, Ben. 2011. @NHS: How the NHS uses Twitter. *The Guardian: Healthcare Network.* Online: *https://www.theguardian.com/healthcare-network/2011/feb/16/nhs-twitter-use-tweets-communication-healthcare*
Whitley, Rob and JiaWei Wang. 2017. Good News? A Longitudinal Analysis of Newspaper Portrayals of Mental Illness in Canada 2005-2015. *The Canadian Journal of Psychiatry* 62. 278-285.
Williams, Raymond. 1976. *Keywords*. Oxford University Press.
WHO. 2009. *WHO: Public Health Campaigns: getting the message across.* Switzerland: WHO Press.
WHO. 2021. WHO Coronavirus (COVID-19) Dashboard. *World Health Organization.* Online: covid19.who.int
Wear, Andrew. 1995. Medicine in Early Modern Europe, 1500-1700. *The Western Medical Tradition, 800 BC to AD 1800,* ed. by Lawrence I. Conrad, Michael Neve, Vivian Nutton, Roy Porter, and Andrew Wear. 215-361. Cambridge University Press.
Wear, Andrew. 2000. *Knowledge and Practice in English Medicine, 1550-1680.* Cambridge, UK: Cambridge University Press.
Weingart, Scott. N.D. Topic Modeling for Humanists: A Guided Tour. Online: scottbott.net.
White, JH. 1935. *The History of the Phlogiston Theory.* London, UK.
Wickham, Hadley. 2010. A layered grammar of graphics. *Journal of Computational and Graphical Statistics.*
Wickham, Hadley, and Garrett Grolemund. 2017. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. 1st ed. O'Reilly Media, Inc.
Wickham et al. 2019. Welcome to the tidyverse. *Journal of Open Source Software,* 4(43), 1686.
Wickham, Hadley, Danielle Navarro, and Thomas Lin Pederson. 2021. ggplot2: Elegant Graphics for Data Analysis. Online: https://ggplot2-book.org/index.html
Wiegand, Viola and Michaela Mahlberg. 2019. Introduction: On context and culture in corpus linguistics. *Corpus Linguistics, Context and Culture,* ed. by Viola Wiegand and Michaela Mahlberg. De Gruyter Publishing.
Williams, Raymond. 1985. *Keywords: a vocabulary of culture and society.* Oxford University Press.
Wootton, David. 2015. *The Invention of Science: A New History of the Scientific Revolution.* Harper Publishing.
Wynne, Martin. 2004. Preface. *Developing Linguistic Corpora: A Guide to Good Practice,* ed. by Martin Wynne. AHDS: Literature, Language and Linguistics.
Yáñez-Bouza, Nuria. 2014. ARCHER past and present (1990-2010). *ICAME Journal 35*.
Yuan, Menglu, Sarah Cross, Sandra Loughlin, and Frances Leslie. 2015. Nicotine and the adolescent brain. *Journal of Physiology. NIH.*
Yun, Gi, David Morin, Sanghee Park, Claire Joao, Brett Labbe, Jongsoo Lim, Sooyoung Lee, Daewon Hyun. 2016. Social media and flu: Media Twitter accounts as agenda setters. *International Journal of Medical Informatics.*

Zacharias, Cody. 2020. twint version 2.1.20: an advanced Twitter scraping and OSINT tool. Online: https://pypi.org/project/twint/
Zarcadoolas, Christina. 2010. The simplicity complex: exploring simplified health messages in a complex world. *Health Promotion International*. Oxford University Press.

APPENDIX

Appendix A: RSC Size of text period divisions

1650 = 2582856

1700= 3414795

1750= 6342489

1800 = 9112274

1850=36993412

1900 = 20159911

Appendix B: US Tokens by article year

Year	**Number of Tokens**
2010	21341
2011	12897
2012	7435
2013	121629
2014	312480
2015	169441
2016	148790
2017	103947
2018	253756
2019	894294
2020	120453

Appendix C: UK Tokens by article year

Year	**Number of Tokens**
2010	1508
2011	4966
2012	18053
2013	94861
2014	273792

2015	187520
2016	158241
2017	113792
2018	137824
2019	288986
2020	31222

Appendix D: US tokens by source

Source	**Number of Tokens**
Arizona Republic	64826
Atlanta Journal Constitution	62322
Baltimore Sun	38911
Boston Globe	102403
Chicago Tribune	131165
Detroit Free Press	37102
Journal Record	14295
Los Angeles Times	173097
NPR	124345
New York Post	49910
New York Times	451087
Newsday	61039
Philadelphia Inquirer	35449
Salt Lake Tribune	53780
South Florida Sun-Sentinel	62950
St. Louis Post-Dispatch	125839
Star Tribune	54467
Texas Tribune	18874
The Wall Street Journal	220650
The Washington Post	244243
USA Today	51720

Appendix E: UK tokens by source

Source	Number of Tokens
BBC	51430
Belfast Telegraph	35307
Daily Mail	307353
Edinburgh Evening News	11099
Mail on Sunday	80491
Manchester Evening News	45774
Sunday Mirror	8281
Telegraph	113315
The Birmingham Post	12310
The Daily Mirror	39711
The Financial Times	72610
The Guardian	95211
The Herald	8057
The London Evening Standard	45700
The Sun	116636
The Times	218567
Wales on Sunday	10626

Appendix F: US Subcorpora Over Time

Years	Tokens
2010-2014	475782
2015-2018	675934
2019-2020	1014747

Appendix G: UK Subcorpora Over Time

Years	Tokens
2010-2014	345482
2015-2017	458606
2018-2020	414988

Appendix H: CDC Accounts

Total corpus size: 1465271

2020: 711218

Tweet Username	**Size**
@cdc_amd	20027
@cdcdiabetes	24225
@cdcemergency	59425
@cdc_genomics	42734
@cdctravel	17238
@cdc_ncezid	64003
@niosh	45645
@fdamedia	38394
@us_fda	91423
@cdcasthma	4761
@cdcdirector	123373
@cdcenvironment	89221
@cdcglobal	34519
@cdcmmwr	40072
@nioshmining	8461
@cdcgov	156510
@nioshconstruct	11842
@veteranshealth	303279
@cdc_cancer	67670
@cdc_ehealth	10552
@cdc_flu	65228
@cdcheart_stroke	10918
@cdc_ncbddd	45083
@niosh_mvsafety	7130
@fda_drug_info	46896
@fdahealthequity	17336
@wtchealthprgm	19306

Subcorpora across 2020	**Size**
January-March	179517
April-June	167272
July-September	187850
October-December	176579

Appendix I: WHO Accounts

Total corpus size: 1274219

Tweet Username	Size
@iarcwho	22416
@ukhcn	737
@whoemro	171394
@whogoarn	4705
@vaccinesafetyn	15392
@lonwho	2838
@whoafro	123227
@ukhcn	737
@whowpro	120378
@pahowho	118648
@whobulletin	7422
@who_europe_vpi	5043
@whosearo	67094
@ukhcn	8461
@who	492380
@who_europe	122545

Subcorpora across 2020	Size
January-March	176533
April-June	244548
July-September	204122
October-December	212224

Appendix J: NHS Accounts

Total corpus size: 2490613

Tweet Username	Size
@thechristienhs	25306
@kingstonhospnhs	34192
@nhsaaa	152905
@nhsdigital	76509
@nhsengland	245478
@nhs_lothian	94014
@nhsnss	41052
@nhsscotland	23623
@northumbrianhs	157722
@publichealthw	78695

@royalfreenhs	96085
@scotgovhealth	25406
@mftnhs	50291
@nhsbt	62989
@nhsemployers	53732
@nhsuk	128162
@nhsmidlands	143008
@nhsnw	49486
@nhsney	71396
@nhsyouthforum	11878
@phe_uk	128832
@primarycarenhs	26675
@kingscollegenhs	33818
@nhs24	54280
@nhscovid19app	142194
@nhsenglandmedia	16676
@nhsenglandldn	103207
@nhsnss	41052
@nhs	238391
@nightingalebham	850
@healthdpt	38091
@ptsafetlynhs	4090

Subcorpora across 2020	**Size**
January-March	297885
April-June	318987
July-September	307395
October-December	442430