Cambridge University Press
978-1-107-12216-1 — Register Variation Online
Douglas Biber , Jesse Egbert
Excerpt
More Information

1 Introduction

1.1 The Expanding Web

It is obvious that computers have fundamentally changed the way we communicate. But the history of that evolution is less well-known. To a large extent, the use of computers as tools for communication can be traced back to the invention of email in the early 1970s. However, computer-mediated communication took decades to become widely available. Given the ubiquitous presence of computers in modern households and our present-day reliance on computers for obtaining information and interacting with others, it is hard to imagine a world where communication was not mediated by computers. However, even fifteen years ago, such forms of communication were relatively rare.

Email communication between two computers was invented in 1972. However, for several years, email communication was restricted to military users connected to the ARPANET. As the ARPANET morphed into the Internet in the early 1980s, it became possible for academic professionals to communicate via email. During this period, university faculty and researchers had widespread access to mainframe computers connected to the Internet, in contrast to the general public, which had much more limited access to computers. However, despite this fact, most academic professionals continued to rely mostly on surface mail for communication well into the 1980s. Instant Messaging (IM) and chatrooms were developed in the 1980s, enabling real-time interactions mediated by computers. These modes of communication were useful for users within an institution (connected to a single mainframe computer), but they had little impact on the ways in which people communicated more generally.

However, in the early 1990s, the rate of change began to accelerate. Personal computers became more widely available, and the World Wide Web became publicly accessible in 1993. At that time, it is estimated that there were only c. 130 websites on the web. But, as Table 1.1 shows, the web has exploded since then, both with respect to the number of users[1] and the number of websites.[2]

Both the number of users and the number of Websites continue to increase right up to the present day. However, the two developments are following

Table 1.1. *Growth of the World Wide Web; 2000–2017*

Year	Number of web users	Number of websites
2000	413,000,000	17,000,000
2005	1,000,000,000	65,000,000
2010	2,000,000,000	207,000,000
2015	3,200,000,000	863,000,000
2017	4,100,000,000	1,767,000,000

Table 1.2. *Growth of the World Wide Web expressed as proportional increases; 2000–2017*

	Percentage increase in the number of:	
Years	web users	websites
2000 to 2005	142%	282%
2005 to 2010	100%	218%
2010 to 2015	60%	317%
2015 to 2017	28%	105%

different trends, as shown in Table 1.2: The number of users more than doubled from 2000 to 2005 and then doubled again from 2005 to 2010. In the last seven years, the number of users continued to increase, but at a much slower rate. In contrast, the number of Websites continued to expand exponentially up until 2015, with an indication of a plateau being reached only in the last two years.

The number of documents found on the web is even more astronomical because most websites contain multiple pages (with most pages including multiple documents). For example, it has been estimated that Google has indexed over 50 billion web pages.[3]

In contrast, the two largest libraries in the world – the British Library and the American Library of Congress – each have around 170 million cataloged items (including books, manuscripts, maps, newspapers, magazines, prints and drawings, music scores, and patents).[4] To put these numbers into perspective, if each document in the British Library was about 1 foot in length, and those documents were all placed end-to-end, they would stretch around the globe once, plus the distance from London to New York. But, if each document on the indexed web was printed out and was about 1 foot in length, and those documents were all placed end-to-end, they would stretch around the globe over 3,800 times!

As a result, any end-user with a personal computer now has direct access to a mind-boggling repository of information, many times larger than the collections in the best libraries of the world. It further turns out that many of the documents collected by major libraries are copyright protected and therefore not available on the public web (see discussion below). Thus, not only is the web many times larger than collections of printed documents, it is also to a large extent nonoverlapping, as the web includes billions of documents that are not available in public libraries.

The web is only a small part of the Internet. The web consists of a system of Internet servers that support documents formatted in HTML (HyperText Markup Language). These HTML documents are linked to other documents, and thus comprise a "web." The Internet includes many additional types of communication that are not part of the public web. By 2000, Internet service providers often offered access to webmail applications as part of their standard packages, making communication through computer networks accessible to the general public. Mobile phone devices were also becoming much more popular and accessible during this same period, and the first text messages were sent in the early 1990s. Social media sites like Myspace and Facebook were first developed in the 2000s and witnessed a remarkable boom in popularity in subsequent years. For example, Facebook was developed for students at Harvard University in 2004, but only five years later (in 2009) it boasted 350 million users. By 2015, this number had increased to more than 1.5 billion users.[5] Many other innovations in online and computer-mediated communication have occurred over the last ten years, such as the development and rise in popularity of Skype and Twitter. As a result, young adults can hardly conceive of a situation where it is not possible to communicate with virtually anyone in the world through multiple channels mediated by computers and technology – and it is even becoming increasingly difficult for older generations to remember such a time.

1.2 The Kinds of Texts Found on the Web

Given these remarkable social and technological changes, it is no surprise that linguists have been intrigued by the possibility of linguistic innovation associated with these new modes of communication. As a result, there have been numerous publications devoted to the study of language on the Internet. Some of the most influential of these include Crystal's 2001 book *Language and the Internet* and Baron's 2008 book *Always On*. There have also been numerous research articles on this topic, and even entire academic journals such as *Language@Internet, The Journal of Internet and Information Systems*, and *The Journal of Computer-Mediated Communication*.

For the most part, these publications have focused on the "special" registers associated with the Internet and computer-mediated communication generally. These are registers that have emerged on the Internet, with no clear counterparts in print media. They include email messages, IM messages, chatroom

Cambridge University Press
978-1-107-12216-1 — Register Variation Online
Douglas Biber , Jesse Egbert
Excerpt
More Information

interactions, interactions in online virtual worlds (MUDs, or Multi-User Dungeons), blogs, discussion forums, social-media postings, etc. These special registers emerged during the 1980s and 1990s, rapidly changing the repertoire of registers available in English (and most other languages around the world). These are the registers that we typically notice when we think of the web, and judging from the coverage of previous books and articles, it would be easy to believe that most of the web consists of these special kinds of texts.

However, even casual surfing on the web quickly indicates that this is not the case. Using a web search engine to identify documents related to almost any topic will return thousands, or even millions, of hits, but most of those documents are not instances of the "special" registers mentioned above. It is not obvious, though, what registers these other documents actually do represent.

Unfortunately, there is no simple way to determine the contents of the web. In Chapter 3, we return to this research issue, applying scientific corpus-based methods to explore the composition of the web. In the present chapter, though, we take a much simpler approach: simply illustrating the types of documents returned by web search engines.

Search engines employ massive databases with indexes of the web documents that are publicly available. It is important to note that the results returned by a search engine do not provide a random sample of web documents. Rather, a search engine employs algorithms that try to prioritize the particular documents that the end-user would most want to see. The normal end-user can employ these search engines to provide a window on the contents of the web. That window is tinted in a way that the search engine chooses: it will look at the particular landscape that the search engine chooses. But the view from that window can still be surprising and not meet our prior expectations of what we expect to see. Most surprisingly, it turns out that "special" web registers are not prevalent in most web searches. Advertisements are also not the most common type of document returned by most web searches. Rather, it appears that the most prevalent registers found on the web are various types of informational documents and news reports.

To illustrate this, we carried out a Google search on the word *horse*, which returned a total of 744 million hits. We coded the first five pages of returned hits (a total of fifty-eight documents), as shown in Table 1.3. The search results summarized in Table 1.3 are typical of many web searches: mostly informational documents, with comparatively few "special" documents and surprisingly few advertisements. Only eight of these fifty-eight documents were from websites for shopping, where the user could buy horses, horse feed, saddles, bridles, or health care services. Surprisingly, even these commercial documents were not overt advertisements. Rather, they mostly presented lists of items/services for sale, with descriptions of the items and the process for purchasing them. In addition, five other documents associated with commercial sites simply presented

Cambridge University Press
978-1-107-12216-1 — Register Variation Online
Douglas Biber , Jesse Egbert
Excerpt
More Information

Table 1.3. *Register categories for the first five pages of hits (fifty-eight documents) returned by a Google search on the word* horse

Register category	Number of documents	% of total
Information about an institution or association	19	33
Informational documents	11	19
News reports	6	10
Commercial sites	8	14
Informational documents associated with commercial sites	5	8
How-to/advice documents	2	4
Blogs	1	2
Discussion forums	1	2
Not related to the animal "horse"	5	8
TOTAL	58	100

information about horses. Three of these documents gave tips for horse breeding, training, nutrition, health care, etc. (*Horse-Journal.com*; *horsechannel.com*; *thehorse.com*), while a fourth document presented extensive information about the Chinese Zodiac Year of the Horse, sponsored by *travelchinaguide.com*.

In addition to the informational documents associated with commercial sites, the search on "horse" returned nineteen informational documents about a horse association or institution (e.g., the American Quarter Horse Association; American Horse Council; Arabian Horse Association; Kentucky Horse Council; Unwanted Horse Coalition; University of Minnesota Horse Program; Luckyorphanshorserescue.org; an association to preserve Idaho wild horses). Another eleven documents simply presented information about horses. Some of these were general encyclopedia articles or dictionary definitions, but others were more specialized (e.g., "horse facts" from National Geographic; "breeds of horses" from Oklahoma State University; research-based information about horse training and health from eXtension.org; a discussion of the origin of horses from quart.us; and information about fossil horses from the Florida Museum of Natural History). News reports can be considered as a more specialized type of informational document. In some cases, the news reports focused on current events (e.g., the outcome of a horse race). However, other news documents in the search provided in-depth discussion of a topic (e.g., the demise of wild horses, or "the ugly truth about horse racing"), making them more informational than narrative reportage. Beyond that, there were a few advice documents (relating to the care or training of horses), one blog posting relating to horses, and one discussion forum.

Of course, it is not possible to evaluate the composition of the web from a single search and only fifty-eight hits. Different words are associated with different aspects of society so we might predict very different types of documents corresponding to those words. But one surprising fact documented in the following chapters is the prevalence of information and news documents, regardless of the particular word or phrase being searched.

For example, we initially wrote this chapter during the Christmas holiday season, when the iPhone 6 was an especially hot item. It is thus reasonable to expect that a web search on "iPhone" would return mostly advertisements or documents associated with commercial sites. Surprisingly, that was not the case. Advertisements and commercial sites accounted for only 10 percent of the total hits for "iPhone." In contrast, news reports accounted for over 50 percent of the documents returned by this search. In addition, there were purely informational documents and numerous reviews, which can be regarded as a special type of informational document that provides a personal evaluation of a product. The overall predominance of informational documents in our search on "iPhone" was similar to what we saw with our search for "horse," with the primary differences being a predominance of news reports and reviews in the case of "iPhone" versus a high proportion of institutional documents and purely informational documents in the case of "horse."

Searching on other terms can result in even higher proportions of news documents and/or purely informational documents. For example, nearly 80 percent of the documents returned by a search on "Syria" (in December 2015) were news reports, including several in-depth discussions of the country or various influential groups of people in the country. Purely informational documents and editorials were also relatively common in this search. In contrast, there was only one blog posting from a university professor and no advertisements or documents related to commercial sites. At the other extreme, a search on the word "electron" returns few news reports but an extremely high proportion of informational documents.

As noted above, using this approach to explore the composition of the web is problematic, because search engines employ algorithms that prioritize documents considered important to the end user. As a result, personal documents (like opinion blogs or discussion-forum advice) might be less likely to appear in the top search results than informational or news documents from major public sources. But those documents certainly do exist on the web. For example, a Google search on "blogs about horses" returned 18.5 million hits, and a Google search on "blogs about Syria" returned fifty-five million hits! Many of these are Personal Blogs, a type of written document that has no direct counterpart in preinternet history. However, a perusal of these web pages shows that a much larger number of them are opinionated informational documents from a news agency or some other institutional site.

1.3 Situating the Searchable Web Relative to Other Discourse Domains

The present book is a description of register variation on the publicly "searchable web": the part of the Internet that all end-users can access with search engines. But there is a very large segment of the web that is not publicly accessible, sometimes referred to as the "deep web." The websites on the deep web are usually password-protected (and sometimes require a fee), associated with institutions, corporations, and publishing companies. For example, institutions, government agencies, and businesses distribute numerous memos, technical reports, and other documents internally to employees on their own networks. Publishing companies offer documents for a fee, including e-books and research articles associated with academic journals. Although it is more difficult to carry out linguistic research on documents in the deep web, personal experience with them indicates that they contain a much greater prevalence of informational documents than the searchable web generally.

In addition, the searchable web does not include the extended Internet used for private communication. Many of the new registers that have been the primary focus of recent linguistic work belong to this domain. These include recently developed registers like Facebook posts and Tweets, as well as registers with a longer history such as email messages and Instant Messages. As documented in books like Crystal (2001) and Baron (2008), these registers arose out of unique communicative circumstances, and as a result, they have developed highly distinctive linguistic characteristics.

The focus of the present book, however, is on those registers that comprise the publicly searchable web. Although they have been generally disregarded in previous linguistic research, these registers have a central place in modern society. A simple reflection of that fact is the rise of the verb *to Google*, referring to the extremely common practice of using the Google search engine to obtain information from the web. Surprisingly, though, we know little about the linguistic characteristics of the registers that result from these searches.

Part of the reason for this neglect is the perception that the informational documents found on the web are the same as informational documents found in print-media, and thus there is nothing new to be learned from a linguistic analysis of those registers. However, this perception is misleading in several respects. In the first place, we simply do not know if web registers are the "same" as print-media registers, until we actually carry out a comprehensive linguistic analysis of web documents. In fact, the descriptions in the following chapters show that this perception is far from accurate. Rather, there are numerous informational web registers that are unlike the print-media registers that we normally encounter. The case study above, on "horse," illustrates this pattern. For example, informational documents associated with a commercial site (see our full discussion in Chapters 6 and 7) are a type of text not normally encountered outside of the web. On the surface, these documents have the primary purpose of conveying information. They are not overtly persuasive, and they certainly do not fit our stereotypes for advertising. However, in many cases, these documents also have an underlying purpose of convincing the reader to make a purchase. This register is relatively pervasive on the web but has no obvious counterpart in the print-media domain.

A related problem, though, is largely methodological rather than a reflection of true differences between the print-media versus web domains. Most linguistic descriptions of print-media registers in the last 30 years have applied the analytical framework of corpus linguistics and have been based on large corpora of texts. The text categories used for the construction of those corpora have had a certain face validity, leading to the perception that linguistic analyses of those corpora fully represent the domain of print-media registers. A more careful reflection, however, quickly reveals that that is far from the case. Most written corpora to date have focused almost exclusively on published texts: novels, academic books and research articles, nonfiction books, magazine articles, and newspapers. In contrast, the population of printed texts that are not officially “published” has been almost entirely overlooked in previous corpora. Those texts include the thousands of informational brochures, reports, and documents found in businesses, medical, professional, and government offices, schools, etc. As a result, previous studies of register variation based on available corpora have described only a part of the entire population of print-media registers.

It turns out that the population of texts available on the searchable web is to a large extent complementary to the population of texts represented in current corpora. As noted above, present-day corpora mostly represent commercially published written texts. In contrast, the searchable web is largely composed of unpublished texts. That is, commercially published texts – like fictional novels, academic research articles, or even many current magazine articles – belong to the domain of the deep web and are not freely available through public web searches. As a result, the sample of written texts available in public libraries and bookstores is mostly non-overlapping with the population of texts available on the searchable web.

Thus, previous descriptions of linguistic variation among written registers, based on available corpora, differ in two major respects from an analysis of web registers: (1) they have focused on print-media registers rather than registers available in an electronic format on the web; and (2) perhaps more importantly, they have focused on traditional published registers (e.g., novels, books, or academic articles) rather than unpublished registers (e.g., informational brochures, instructional pamphlets, product reviews, or personal letters). The present book, by focusing on the full range of documents found on the searchable web, provides a first step toward filling this gap.

Finally, the study of discourse from the searchable web is theoretically important because it causes us to rethink traditional notions of "register." Since the 1960s, written corpora have been organized in terms of major textual categories, which we refer to as "registers." Those categories have been treated as if they are relatively uncontroversial: Published print-media texts usually have overt external indications of register, and thus it has not proven difficult to classify individual texts. For example, newspaper articles are published in newspapers; magazine articles are published in magazines; academic articles are published in academic research journals; novels are published as books and explicitly claim to be fictional; etc. Even specific registers often have external indicators. For example, news reportage articles are published on the front page of a newspaper (and in the "International" and "National" sections of the newspaper); sports reports are printed in the "Sports" section of the newspaper; editorials and letters to the editor are published on the editorial pages of the newspaper. These external criteria are usually sufficient for classifying written texts into register categories, and, as a result, it has not been considered to be problematic for discourse analysts (and corpus compilers) to identify the register of individual texts.

In contrast, the documents returned by a web search often have little or no indication of register category. For example, the Google search on "horses" described above returned millions of documents. Some of these documents have external indicators that help to identify their register, such as an encyclopedia article from Wikipedia, a newspaper story from the *New York Times*, or a magazine article from the *Atlantic*. However, the register category of many other documents is more nebulous, such as: an informational page about horses from the *Oklahoma State University Horse Project*; a page giving "Fun horse facts for kids" from *Sciencekids.co.nz*; a short informational text about horses from *PBS*; a guide to equine health care from *thehorse.com*; and descriptions of horse associations (e.g., the *Arabian Horse Association*, the *American Paint Horse Association*). Such web documents are familiar to any end-user of the web. But unlike most published print-media texts, the register category of these web documents is not obvious.

Observations like these lead to one of the central themes of the linguistic descriptions in the present book: that most web documents are not "pure" instances of a particular register, and that even the register categories themselves might sometimes be understood as "hybrids" that serve multiple communicative purposes (combining narrative, informational, opinionated, and how-to/advice purposes in different ways and to differing extents). By extension, these observations raise general theoretical questions about the categorizations used in previous corpus-based studies of print-media registers, raising the possibility that a hybrid perspective applied to the domain of print-media registers might also be productive. While such analyses are well beyond the scope of the present book, they do raise interesting theoretical questions regarding the notion of "register." We thus return to these issues in the concluding chapter of the book.

1.4 Overview of the Book

As noted in the sections above, the present book is innovative in three key respects:

1. It focuses on analysis of the full range of registers found on the public searchable web, rather than being restricted to a description of a few specialized Internet registers (like Tweets, Facebook posts, etc.).
2. It focuses on freely available written registers, which can be considered unpublished in the traditional sense. This focus is in marked contrast to previous corpus-based studies of written registers in the print-media domain, which have focused almost exclusively on commercially published written registers.
3. It recognizes the existence of hybrid registers, and undertakes analyses that explore different ways in which web registers are hybrid, with respect to both their situational characteristics and their linguistic characteristics.

Our linguistic descriptions are empirical, based on analysis of a large corpus of web documents: a near random sample of c. 48,000 documents from across the entire spectrum of the publicly searchable web. Chapter 2 of the book describes our methods for constructing and coding this corpus. In the initial stages of the project, the corpus consisted simply of web documents, with no indication of the register categories for those documents. In fact, we began with no preconceptions of what those register categories would be. Then, using crowdsourcing techniques, with ratings from actual end-users of the web, we developed a taxonomy of online register categories, coding each document in our corpus for its register. The taxonomy and the process of coding are described in detail in Chapter 2.

In Chapter 3, we describe the register composition of our corpus as an indication of the composition of the searchable web more generally. Eight general registers are distinguished, with numerous specific sub-registers within the general categories. We describe the relative frequency of each register category in our corpus and further introduce the possibility of hybrid registers on the web.

In Chapter 4, we move on to the overall linguistic description of the patterns of register variation, applying Multidimensional analysis. The chapter begins with an overview of the methodological framework of Multidimensional analysis, and then describes the nine linguistic "dimensions" that emerged in that analysis, coupled with descriptions of the similarities and differences among web registers with respect to each dimension.

Building on that foundation, Chapters 5–8 provide more detailed linguistic descriptions of the major registers found on the searchable web: narrative web registers (Chapter 5); opinion, advice, and persuasion web registers (Chapter 6); informational descriptions, explanations, and procedures (Chapter 7); and oral web registers (Chapter 8). These chapters document the range of specific sub-registers within each of these general categories, and describe the distinctive situational, grammatical, and lexical characteristics of those sub-registers.

Finally, Chapter 9 concludes the book with a synthesis of our research findings, a description of ongoing and future research in this area, and a discussion of the theoretical implications of this research for studies of register variation in other discourse domains. In particular, we take up the theoretical issue of how register can be investigated in a continuous space of variation. The study of registers on the searchable web forces such a perspective, but we argue in Chapter 9 that this perspective might be equally informative for the study of registers in other discourse domains. Thus, it is our hope that the present book will prove useful both for its detailed linguistic descriptions of web registers as well as its theoretical discussions of issues relating to the discourse construct of “register.”

Notes

1. www.internetlivestats.com/total-number-of-websites/.
2. www.pewinternet.org/2014/03/11/world-wide-web-timeline/; https://www.internetworldstats.com/stats.htm.
3. https://google.com/insidesearch/howsearchworks/thestory/; www.worldwidewebsize.com/.
4. https://en.wikipedia.org/wiki/List_of_largest_libraries.
5. www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/.