*1 Exploring the composition of the searchable web: a corpus-based taxonomy of web registers

Douglas Biber,[2] Jesse Egbert[3]
and Mark Davies[4]

Abstract

One major challenge for Web-As-Corpus research is that a typical Web search provides little information about the register of the documents that are searched. Previous research has attempted to address this problem (e.g., through the Automatic Genre Identification initiative), but with only limited success. As a result, we currently know surprisingly little about the distribution of registers on the web.

In this study, we tackle this problem through a bottom-up user-based investigation of a large, representative corpus of web documents. We base our investigation on a much larger corpus than those used in previous research (48,571 web documents), and obtained through random sampling from across the full range of documents that are publically available on the searchable web. Instead of relying on individual expert coders, we recruit typical end-users of the Web for register coding, with each document in the corpus coded by four different raters. End-users identify basic situational characteristics of each web document, coded in a hierarchical manner. Those situational characteristics lead to general register categories, which eventually lead to lists of specific sub-registers. By working through a hierarchical decision tree, users are able to identify the register category of most Internet texts with a high degree of reliability.

After summarising our methodological approach, this paper documents the register composition of the searchable web. Narrative registers are found to be the most prevalent, while Opinion and Informational Description/Explanation registers are also found to be extremely common. One of the major innovations of the approach adopted here is that it permits an empirical identification of 'hybrid' documents, which integrate characteristics from multiple general register categories (e.g., opinionated-narrative). These patterns are described and illustrated through sample Internet documents.

[2] Doug Biber Douglas.Biber@nau.edu
Jesse Egbert jesse_egbert@byu.edu
Mark Davies (No email provided)
[3]
[4]

Keywords:

1. Introduction

There is a mind-boggling amount of information available on the World Wide Web. For example, Fletcher (2012: 1) estimates that Google indexes about 40 billion web pages. And it is similarly mind-boggling to consider how many people rely on the Web as a source of information. In June 2012, there were an estimated 2.5 billion users of the Web worldwide, with more than half of those individuals using the Web every day.[5] It is difficult to obtain exact current information on Internet use, but the Web has clearly become important as a source of information in everyday life.

Despite this important role, the Web is a mysterious black box for most end-users: a user enters a term into a search engine and the search engine returns links to web pages. But there is little understanding on the part of the user about the population of documents that was searched or what the search procedure entails. Many users have become adept at modifying their search methods to obtain the information that they want. Thus, although they do not actually know the contents of the black box, most users do not experience dissatisfaction with the information returned by Web searches.

In contrast, the mystifying composition of the Web can be more problematic for linguists using the web as a corpus to investigate patterns of language use. This approach has become so prevalent that the acronym WAC (Web-as-Corpus) is now commonplace among researchers who explore ways to mine the WWW for linguistic analysis. One of the major challenges for WAC research is that a typical Web search usually provides us with no information about the kinds of texts that are gathered. For example, Fletcher (2012: 1341) notes that a linguistic search of the Web-as-Corpus will tell us nothing about:

> For whom and what purpose is the text intended? What […] target audience does it represent? Was it written carefully or carelessly by a native speaker, or is it an unreliable translation by man or machine? Is the document authoritative – accurate in content and representative in linguistic form?

Similar problems were noted a decade earlier by Kilgarriff and Grefenstette (2003) in their introduction to the special issue of the journal Computational Linguistics on WAC. They write:

---

[5] See: http://www.internetworldstats.com/stats.htm and http://www.thecultureist.com/2013/05/09/how-many-people-use-the-internet-more-than-2-billion-infographic/

> 'Text type' is an area in which our understanding is, as yet, very limited. Although further work is required irrespective of the Web, the use of the Web forces the issue. Where researchers use established corpora, such as Brown, the BNC, or the Penn Treebank, researchers and readers are willing to accept the corpus name as a label for the type of text occurring in it without asking critical questions. Once we move to the Web as a source of data, and our corpora have names like 'April03-sample77,' the issue of how the text type(s) can be characterized demands attention.

These concerns are shared widely among WAC researchers and, as a result, there has been a surge of interest over the last few years in Automatic Genre Identification (AGI): computational methods using a wide range of descriptors to classify web texts by genre (or register) categories automatically. Computational linguists in this research area have used the cover term 'genre' for the text categories on the Web. However, in this paper, we employ the term 'register' rather than 'genre' to refer to situationally based textual distinctions on the Web, following the research tradition developed in Biber (1995), Biber *et al*. (1999), and Biber and Conrad (2009).[6]

Of course, the prerequisite for computational techniques that automatically identify the register of a web document is a taxonomy of the possible register categories found on the web. That is, it is not possible to develop and test methods for the automatic prediction of register until we know the full set of possible web registers. In addition, it would be beneficial to know the distribution of those registers: which ones are especially prevalent, and which ones are rare. To date, however, efforts to obtain this information have had limited success.

One major problem in this regard is that web documents often have few, if any, external indication of register category. In contrast, written/published texts usually have overt external indications of register. For example, newspaper articles are published in newspapers; magazine articles are published in magazines; academic articles are published in academic research journals; recipes are published in cookbooks; and personal diaries are written in some kind of a personal journal. Even specific registers often have external indicators. For example, news 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

[6] Biber and Conrad (2009) develop an analytical framework that distinguishes between 'registers' and 'genres' (see especially pages 15–23). In brief, registers are situationally defined varieties. Specific linguistic features are pervasive and frequent in a register because they have a systematic functional association with the situational context. Genres are also associated with particular situations of use, but the linguistic characteristics associated with the genre perspective are conventional rather than functional. End users of the web can pay attention to both register and genre characteristics, so the distinction is somewhat blurred for the perceptual categorisation of texts.

'sports' section of the newspaper; and 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 so, as a result, it is not problematic for discourse analysts (and corpus compilers) to identify the register of individual texts.

In contrast, the web documents returned by a web search often have little or no indication of register category. For example, a simple Google search on 'horses' returns hundreds of pages. Many of these have identifiable registers, such as an encyclopedia article from Wikipedia, a newspaper story from the *New York Times*, and a magazine article from *The Atlantic*. However, the register category of many others 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 written texts, the register category of these documents is not obvious.

Corpus-based analyses of written/published texts also differ from corpus analyses of Web documents in their scope: the population of documents found on the web is much more diverse than the population normally represented in a corpus of written (published) texts. That is, corpora are normally sampled from published written sources: newspapers, magazines, books, *etc*. There are some corpora that are designed to represent unpublished student writing, and there are also a number of corpora sampled from spoken domains. However, few corpus researchers have attempted to collect a representative sample of unpublished written texts, such as junk mail, posters and flyers, personal letters, course handouts, how-to instructions, *etc*. And even fewer researchers have attempted to classify such texts into register categories, or assess whether their sampling of such texts is representative. As a result, published written texts have had a privileged status in much previous corpus research, while unpublished written texts have been generally disregarded.

In contrast, unpublished documents have equal status with published texts on the Web, and, in fact, it is often difficult to distinguish between the two. As a result, any taxonomy of the register categories on the Web must include both the textual distinctions used in standard written corpora (e.g., novels, academic research articles and newspaper editorials) as well as a slew of other kinds of text that do not find their way into published sources (e.g., a personal information page about rock collecting; a personal page about travelling in Thailand; and a personal page with opinions about the quality of rock climbs). The challenges for AGI have, thus, been to develop a taxonomy of the possible registers found on the web and to develop methods that allow analysts to determine the register category of individual documents reliably – as prerequisites to

computational techniques that can be used to identify register categories automatically.

The typical methodology used in an AGI study is to begin with a manual coding of genre/register for the documents in a limited sample from the Web, and then to test the extent to which computer programs can automatically place those texts into the same categories. However, although some studies have achieved high accuracy rates (e.g., Lindeman and Littig, 2010; and Santini, 2010), serious questions have been raised about the validity of those results. First, some scholars raise doubts about the representativeness of the web corpora that are analysed in those AGI studies. These corpora are typically small (around 1,000 web documents) and often not sampled in a way that ensures representativeness (see the discussion in Santini and Sharoff, 2009: 131–3).

A second issue concerns the methods used to code documents. Most studies begin with a list of possible register categories, and then Internet documents are manually classified into those categories by an 'expert' – typically the primary researcher. This approach is based on the assumption that a single expert user is able to 'correctly' identify the register category of individual Internet texts. Unfortunately, that assumption does not seem warranted: in the few cases where inter-rater reliability has been evaluated, it is reported to be quite low (even among expert linguists as raters). This is especially true for corpora comprised of randomly extracted web texts (see discussion in Sharoff *et al*., 2010: Section 3.4). Given the problems that 'experts' have in identifying web genre categories, it is not surprising that non-expert web users also vary in their understanding of genre labels (see Crowston *et al*., 2010), and that reliability among lay users is often unacceptably low (Rosso and Haas, 2010).

Most importantly, though, there has been no agreement to date concerning the 'correct' set of possible register categories found on the Web. That is, AGI researchers usually begin with a set of possible genre/register categories based on an *a priori* intuitive consideration of Internet texts. In practice, each researcher employs a different set of categories, which can vary widely. This problem has been recognised by and been discussed in previous research; thus, for example, Rehm *et al*. (2008: 352) note:

> One of the most important problems concerns the elusiveness of the concept of genre. The consequence is that, in practical terms, genre researchers usually have different ideas of what a genre is, how genres should be defined and identified and, therefore, they use different genre labels in their approaches.

A few years ago, there was a considerable effort to agree on a standard set of register/genre categories for AGI research, as part of a wiki-

based collaboration among web-as-corpus experts.[7] That collaborative effort resulted in a list of seventy-eight register/genre distinctions; but the initiative appears to have faded out in the last few years, with little consensus regarding the relative status of those categories. As a result, there is still no set of register/genre categories used in current AGI research that has been generally agreed on.

In this study, we tackle these problems with an alternative approach. First, we base our investigation on a much more representative corpus – much larger than in previous research (48,571 web documents) – obtained through random sampling from across the full range of documents that are publically available on the web (see Section 2.1). Second, instead of relying on individual expert coders, we recruit end-users of the Web for our register coding, with each document coded by four different raters; this allows us to assess the degree of agreement among users. Finally, and most importantly, we do not force users to choose directly from a pre-defined set of specific register categories. Rather, we ask users to identify basic situational characteristics of each web document, coded in a hierarchical manner (see Section 3). Those situational characteristics lead to general register categories, which eventually lead to lists of specific sub-registers. By working through a hierarchical decision tree, users are able to identify the register category of most Internet texts with a high degree of reliability.

In Section 2, we briefly document the methodological procedures used for this project. (Readers are referred to Egbert *et al*., 2014, for a more detailed discussion of the methods.) In Section 3, we introduce the register framework used for our study. In the main body of this paper, then, we describe our findings. We begin in Section 4 with distributional results, describing the overall prevalence of different types of registers on the web. These quantitative results also led us to include a more specialised type of register identified by users: 'hybrid registers'. Then, in Section 5, we describe the communicative and textual characteristics of these registers in more detail, with a special focus on the ways in which hybrid registers integrate characteristics from multiple general register categories (Section 5.6).

2. Methods

2.1 Corpus for analysis

The corpus used for our study was extracted from the 'General' component of the Corpus of Global Web-based English (GloWbE).[8] The GloWbE corpus contains about 1.9 billion words in 1.8 million web documents, collected in November to December 2012 by using the results of Google

[7] See: http://www.webgenrewiki.org/
[8] See: http://corpus2.byu.edu/glowbe/

searches of highly frequent English 3-grams (i.e., the most common 3-grams occurring in COCA; examples include 'is not the' and 'and from the'). We saved between 800 and 1,000 links for each n-gram (i.e., 80–100 Google results pages), thereby minimising the bias from the preferences built into Google searches. Many previous web-as-corpus studies have used similar methods with n-grams as search engine seeds (see, for example, Baroni and Bernardini, 2004; Baroni *et al*., 2009; and Sharoff, 2005, 2006). It is important to acknowledge that no Google search is truly random. Thus, even searches on 3-grams consisting of function words (e.g., 'is not the') will to some extent be processed based on choices and predictions built into the Google search engine. However, selecting hundreds of documents for each of these n-grams that consist of function words rather than content words minimises that influence.

To create a representative sample of web pages for us to analyse in our project, we randomly extracted 53,424 URLs from the GloWbE Corpus. This sample, comprising web pages from five geographic regions (United States, United Kingdom, Canada, Australia and New Zealand), represents a large sample of web documents collected from the full spectrum of the searchable Web. Given that the ultimate objective of our project is to describe the lexico-grammatical characteristics of web documents (see Section 6), any page with less than seventy-five words of text was excluded from this sample.

To create the actual corpus of documents used for our study, we downloaded the web documents associated with those URLs using HTTrack.[9] However, because there was a seven-month gap between the initial identification of URLs and the actual downloading of documents, about 8 percent of the documents (n=3,713) were no longer available (i.e., they were linked to websites that no longer existed). This high attrition rate reflects the extremely dynamic nature of the universe of texts on the Web.

Our ultimate goal in the project is to carry out linguistic analyses of Internet texts from the range of web registers (see our discussion in the Conclusion). For this reason, 1,140 URLs were excluded from subsequent analysis because they consisted mostly of photos or graphics. Thus, the final corpus for our project contained 48,571 documents. To prepare the corpus for POS tagging and linguistic analyses, non-textual material was removed from all web pages (HTML scrubbing and boilerplate removal) using JusText.[10]

2.2 Overview of procedures

The study described here marks the first stage of a major project to undertake comprehensive linguistic analyses of the patterns of register

[9] See: http://www.httrack.com
[10] See: http://code.google.com/p/justext

variation on the Web, as the basis for automatic register (genre) identification. As a prerequisite for those goals, it was necessary, first, to identify the registers found on the web and document the extent to which each of those registers is actually used. This paper reports on that stage of the research.

Identifying the set of registers found on the web proved to be a challenging task. Our goal here was to establish a set of register distinctions that end-users actually recognise and can reliably identify. We tested several different approaches for this task, and eventually decided to use a decision tree of situational characteristics that lead to specific registers, rather than asking users to identify directly the register category of a given Internet document. We describe this approach and the associated register distinctions in Section 3.

We developed a computational tool for register classification, which we implemented on Mechanical Turk (an Amazon-based online crowd-sourcing utility). We tested this tool through several rounds of piloting, resulting in numerous revisions to both the tool and general approach. This development process, and the preliminary results of the pilot testing, are documented in Egbert and Biber (2013) and Biber and Egbert (2013).

Given the encouraging results of our final pilot studies, we proceeded to the analysis of our full corpus of 53,424 web documents. We recruited 908 raters through Mechanical Turk for this task. Before a rater was allowed to participate in the task, they were required to complete a seven-minute interactive tutorial video and code a practice document. Responses were checked for quality before raters were approved. Each rater was paid $0.11 for each document that they classified and each document was classified by four independent raters. Two major steps were taken to ensure quality control during the rating process: first, we randomly spotchecked a sample of early ratings and occasionally contacted raters to give them further training. Then, throughout the entire rating process, we frequently reviewed results, sorting them by rater, register category and time spent per rating, in order to check for questionable patterns and unsatisfactory work. In general, we were amazed by the high level of enthusiasm and care that raters exercised during the coding process. There was extensive forum discussion among raters during the entire process, and we received hundreds of e-mail messages from raters asking for guidance and feedback. It would not have been feasible to code a corpus of this size (with four raters per document) without the aid of a resource like Mechanical Turk. However, beyond that, we felt that the quality of coding that we received from MTurk raters exceeded what we would normally have received from student participants. (See Appendix A for a sample of some of the e-mails we received.)

As noted above, 3,713 of the web pages in our initial sample were no longer available, and our raters coded another 1,140 of these pages as consisting mostly of photos or images. Thus, the final corpus for analysis

consists of 48,571 documents, with each document coded for register categories by four independent raters.

3. Register categories distinguished in the study

Before undertaking empirical investigation of the registers found on the web, we needed to compile a representative corpus where each document was coded for its register category. We initially asked non-expert users of the Internet to identify the register category of individual web pages directly (see the detailed description of our pilot research in Egbert *et al*., 2014). However, for the reasons described in Section 1, that approach proved to be unsuccessful, with agreement rates below 50 percent in some cases.

As a result, we developed an alternative approach, asking users to code basic situational characteristics rather than directly coding a register category. Building on the situational framework for register description developed in Biber and Conrad (2009: Chapter 2), we asked raters to code the mode (spoken or written), relations among participants (multiple interacting participants *versus* authors who do not interact with addressees), and communicative purposes (e.g., to narrate, to inform, to express opinion). Then, based on those choices, we offered users a set of specific registers to choose from under the more general register characteristics. Table 1 provides an overview of the coding framework.

The framework is organised as a hierarchical decision tree, with each level representing a different situational parameter. At the top level, we asked users to make a two-way decision about the mode of production:

- Internet texts that originated in the spoken mode (e.g., transcripts of speeches or interviews)
  versus
- Internet texts that originated in the written mode

For the written texts, we then asked users to distinguish between interactive discussions among multiple participants (e.g., discussion forums) *versus* non-interactive Internet texts. Even this simple distinction is often not clear-cut on the web, because authored web documents are often followed by reader comments. We thus made it clear to coders that 'written interactive discussions' are distinct from written documents followed by reader comments, and that coders would be able to note the existence of reader comments for non-interactive texts later in the process.

For non-interactive written texts, we asked users to distinguish among six general register categories based on communicative purpose – that is, to:

- Narrate or report on EVENTS [past, present, or] (news report/blog, sports report, personal/diary blog, historical article, short story, novel, biographical story/history, magazine article, travel blog, *etc*.);
- Describe or explain INFORMATION (description of a person, description of place/product/organisation, FAQs about information, research article, informational blog, technical report, legal terms and conditions, *etc*.);
- Express OPINION (opinion blog, review, advice, advertisement, religious blog, letter to the editor, self-help, *etc*.);
- Describe or explain FACTS WITH INTENT TO PERSUADE (editorial, description with intent to sell, persuasive article or essay, *etc*.);
- Explain HOW-TO or INSTRUCTIONS (how-to, instructions, FAQ, recipes, technical support, *etc*.); and,
- Express oneself through LYRICS (song lyrics, poem, prayer, *etc*.)

Finally, after a user had identified these general register characteristics, we asked them to select a specific sub-register under the general category. For example, 'interviews' and 'TV scripts' were possible choices under the 'Spoken' general category, while news reports and travel blogs were possible choices of specific registers under the written-noninteractive-narrative general category (see Table 1).

==Insert Table 1 about here==

4. Agreement results for the coding of registers and sub-registers

Overall, raters were able to achieve 'moderate agreement' for their coding of the general register category of the 48,571 documents in our corpus (Fleiss' Kappa = 0.47).[11] A more detailed consideration of the levels of agreement, however, shows stronger results and allows us to offer an interpretation of the register category for most documents in our corpus. Thus, Table 2 shows that raters were able to achieve majority agreement (at least three of the four raters) on the general register category for almost 70 percent of the web pages in our corpus. All four raters agreed on the classification of about 37 percent of the texts, while three of the four raters agreed on the classification of an additional 32 percent (approximately) of the texts. For 29.2 percent of the documents, raters had a split involving a combination of two (or three) registers. It was established, though, that a

[11] Fleiss' Kappa is a measure of inter-rater agreement that is well-suited to the design of our study. Unlike related measures of inter-rater agreement (e.g., Cohen's Kappa), Fleiss' Kappa does not require that the same raters code each of the documents in the dataset. However, like Cohen's Kappa, Fleiss' Kappa accounts for chance agreement among raters, making it more robust than simple percent agreement. Kappa values in the range of 0.41–0.60 can be interpreted as 'moderate agreement' (Landis and Koch, 1977).

few of the specific combinations in these splits occurred repeatedly in the corpus. As a result, in Section 5.6 we explore the possibility that these common 2–2 and 2–1–1 splits represent interpretable 'hybrid registers'. Overall, these results show that non-expert web users can, to a large extent, meaningfully classify web pages into general register categories.

==Insert Table 2 about here==

The levels of agreement were lower for the coding of specific sub-register categories (Fleiss' Kappa = 0.40). Table 3 shows that raters were able to agree on the sub-register for about 51 percent of the web pages (with three or all four raters in agreement), but there was no agreement at all on the specific sub-register for 11.3 percent of the documents. Further, many 2–2 and 2–1–1 hybrid combinations at the sub-register level are not systematic and not easily interpretable (see Section 5.7).

==Insert Table 3 about here==

These results reflect the difficulty of identifying specific sub-register categories for web documents (see discussion in Section 1), and the usefulness of our hierarchical approach based on simple situational characteristics and communicative purposes. In general, raters were able to agree on those situational characteristics and the associated general register categories, but they experienced considerable difficulty in determining the specific sub-register. For example, many documents were classified as 'non-interactive written informational description / explanation' by all four raters. But those same raters were often unable to agree on specific sub-registers, so that the same document might be classified as an informational explanation, informational blog, a description of a person, informational FAQs, legal terms and conditions or an encyclopedia article. As we show below, this difficulty does not apply uniformly to all sub-registers. Rather, some documents are readily categorised, while there is almost no agreement on the specific categories of other documents.

5. Exploring the composition of the web

The data obtained from the end-user coding process (described in Sections 2 to 4) allows us to explore the composition of the web, asking what registers are especially prevalent and which ones are relatively rare. Thus, Table 4 shows the breakdown of general register categories (presented in order of frequency) for the 33,619 documents that raters agreed on (i.e., the documents where three or four raters were in agreement; see Table 2).

==Insert Table 4 about here==

Our perceptions regarding the composition of the web are coloured by the searches that we typically do, and by the pages that search engines direct us to. As a result, few of us have accurate intuitions concerning the actual composition of the web. Based on most users' experience, we might predict that advertisements are the most frequently found type of document on the web. Table 4 shows, however, that this is not the case. Informational Persuasion documents are usually a kind of indirect advertisement, presenting descriptive information about a place or product with the goal of persuading the reader to purchase something. However, those documents are not prevalent in our random sample of web pages (only about 1.6 percent of the total), and, otherwise, there are few overt advertisements in our corpus.

The scarcity of advertisements in our corpus can be attributed to several factors. First, typical usage of the Internet can lead a user to believe that advertisements are more prevalent than they actually are. Many users commonly shop online and thus regularly encounter advertisements on commercial sites, and search engines are structured to direct users to commercial sites, even when users are not shopping online. A second major factor, though, is that many advertisements on the web are not (primarily) textual and are not, thus, represented in our corpus. For example, pop-up web pages are not part of the searchable web (and so are not included in our sample), and advertisements found on the sides of a web page were removed in the 'scrubbing' process of our corpus creation. In addition, our corpus includes only pages with at least seventy-five words of prose, and so excludes all advertisement pages consisting mostly of photos with little prose. However, even considering all of these factors, the results presented in Table 4 show that our perceptions of the web can be dramatically different from its actual composition, and, in particular, that advertisements do not dominate the textual content of the searchable web.

Instead, our findings show that narrative texts are by far the most common register on the web: 31.2 percent of all documents in our corpus. In addition, a large proportion of the hybrid documents include narrative purposes (see discussion in Section 5.6). As a result, over 50 percent of all documents on the web have a narrative purpose. Furthermore, Informational Description/Explanation documents (about 14 percent of the corpus), and opinion documents (about 11 percent of the corpus) are both prevalent on the Web; and the majority of hybrid documents (Section 5.6) also include one or both of these communicative purposes.

Thus, taken together, the three general register categories of Narrative, Informational Description/Explanation, and Opinion account for well over 80 percent of the documents on the Web. In contrast, the other five categories (Interactive Discussion, How-to/Instructional, Informational Persuasion, Lyrical and Spoken) are considerably less common. In the following sections, we provide more detailed descriptions of each of these general registers.

5.1 Narrative

Table 5 shows that half (52.5 percent) of the narrative texts in our corpus are general news reports, while an additional 16 percent are sports news reports. Many of these texts are examples of registers found in print media that have simply been transferred to the web. Others are news reports that have been incorporated into a regular blog. At first, we planned to distinguish news blogs from regular news reports (which have their origins in print media). In practice, though, it proved nearly impossible to determine whether a news/sports report was originally published in a print newspaper or whether it had been written specifically for a blog. As a result, we combined reports and blogs to form a single category. We did, however, distinguish sports reports/blogs as a specialised sub-category of general news reports/blogs.

==Insert Table 5 about here==

Taken together, news reports and sports reports comprise 21.4 percent of the entire corpus (i.e., 10,411 of the 48,571 documents in the corpus). This percentage is considerably higher when we include hybrid texts that can be treated as news reports combined with some other purpose (e.g., news blogs that report on events with an opinionated bias; see Section 5). Thus, well over 25 percent of the searchable Web consists of news reports with a narrative focus, packaged in many different ways, from a bewildering array of sources and focussed on an incredible range of topics (past events involving nations, sports teams, celebrities, entertaining stories, *etc*.).

The personal narrative blog – recounts of past personal events – is also an important sub-register in this category, comprising 11.3 percent of all narrative web documents, or about 3.5 percent of all documents in our corpus. From a technical perspective, it is difficult to formulate a precise operational definition of 'blogs'. However, end-users of the web have little trouble identifying many web documents as clear instances of blogs because they are often explicitly labelled as 'blogs'. There are many specific sub-types of blogs (see Herring *et al*., 2005; and Sindoni, 2013: Chapter 3); in our framework, we included blog sub-registers under the general register categories of Narrative (news reports/blogs, sports reports/blogs, personal narrative blogs, travel blogs), Informational Description (informational blogs), and Opinion (opinion blogs).

The blogs grouped into the category of personal narratives are recounts of past events that the blogger participated in; for example, daily experiences while learning how to weave, events that occurred with a new baby, experiences at a fashion week and experiences being outdoors during the winter. Although we treat them as a separate category in our analysis,

travel blogs can also be regarded as a special sub-type of personal narrative blog. It is not clear who the intended audience is for many blogs or how widely read they are. But there is no shortage of authors – people who are eager to share their own personal experiences and opinions with a public audience (see Guadagno *et al*., 2008; and Sindoni 2013: Chapter 3). The high rate of occurrence of these personal narrative blogs is especially remarkable given that our corpus is sampled from the searchable public web, and thus excludes private social media messages, where there is presumably much more of this type of communication (see Sindoni 2013: 120–3).

All other sub-registers of narrative are considerably less frequent on the web, including many registers that are widely recognised in print media (e.g., historical articles, short stories, novels and biographies). This illustrates a general trend emerging from our study – that the most common registers found on the web are not those typically analysed in corpora of published written texts. Conversely, although the most widely analysed registers from published writing can be found on the web, they are typically rare in comparison to other web registers.

5.2 Informational Description/Explanation

The second most frequent general register on the web is Informational Description/Explanation (14.5 percent of the total documents in our corpus; see Table 4). This category includes the informational registers that are typically analysed in corpora of published written texts, such as research articles, abstracts, encyclopedia articles and technical reports. However, in common with the pattern observed for narratives, these sub-registers from print media are generally rare on the web in comparison to other types of texts. For example, academic research articles – the focus of an extensive body of corpus-based research – comprise less than 3 percent of the general 'informational' register.

Encyclopedia articles are a special case here: they are not especially prevalent or important in published media, but they are prominent on the Web, with links to encyclopedia articles being returned by many searches. In the every-day experience of a typical college student or teacher, Internet encyclopedia sites are hugely important, having become a source of information on many topics. For this reason, it would be easy to assume that a large part of the Web contains encyclopedia articles. However, Table 6 shows that this is not the case: encyclopedia articles comprise only 6.6 percent of the general informational register category, which corresponds to less than 1 percent of our total corpus.

Table 6 shows that most of the documents in the general informational register are not instances of specific well-defined sub-registers/genres. When we were developing the register classification framework, we had difficulty identifying other named sub-registers in this

category, and so, as a result, we included categories associated with communicative purposes (simple description, description of a person) and specific format (informational blog). The most important of these sub-registers is simple description, which is specified as including descriptions of a place, product, organisation, program, job, *etc*. Descriptions of a person are a related sub-register of this category. Informational blogs are very similar in purpose, but are distinguished primarily by their format. Taken together, these three sub-registers comprise about 30 percent of the documents in the Informational category. These are mostly non-institutional documents presenting descriptive information about almost any conceivable object or topic. Many of these are descriptions of tangible objects or places, such as hotels, restaurants, towns, national parks, types of gems and minerals, types of bolts and screws, useful tools for gardening, *etc*. Some other documents in these categories provide information about more abstract processes or concepts, such as statistics about different countries around the world, a description of the Office of the Director of Public Prosecutions, and a description of the Sideloader Delivery method.

==Insert Table 6 about here==

Most of the documents in this general register did not fit tidily into any of these specific sub-registers. That is, although raters had no difficulty agreeing that these documents were instances of the general category Informational Description/Explanation, they were unable to agree on a specific sub-register for 53.9 percent of the documents. These are mostly informational documents prepared by various organisations, government entities, and other institutions, describing and explaining information related to almost any conceivable topic. They tend to be more technical, abstract and conceptual than the documents that raters classified as simple description, but the framework failed to provide specific sub-register categories that clearly fit the purposes of these documents. Some examples include documents about:

- The advantage of parallel circuits over series circuits;
- Food safety following floods;
- Stress can become a serious illness;
- Attention Hunters: 'It's time to Get the Lead Out' (an announcement prohibiting lead bullets);
- The Major Planets in October 2011;
- Middle schooling – Rationale; and,
- What is a trustee?

These documents are all clearly informational – some more descriptive, and others more explanatory. On the whole, they tend to be technical in content, although they are often packaged for a general readership. Raters had no

trouble identifying these as instances of a general Informational Description/Explanation register. Documents like these illustrate the utility of our hierarchical approach, where it is easy for raters to identify basic situational parameters but nearly impossible to agree on specific sub-register categories.

5.3 Opinion

Opinion web pages are nearly as common as informational pages (see Table 4). As Table 7 shows, a third of these were classified as Opinion blogs (37.9 percent), while another 21 percent were classified as Reviews. As per personal narrative blogs, it is not clear who writes opinion blogs and how many people read most of these blogs. But it is clear that both of these forms of expression are popular. The difference between the two types of blogs concerns the primary communicative purpose: narrating past events *versus* expressing opinions about government, society, *etc*.

==Insert Table 7 about here==

Reviews differ from opinion blogs in that they have a specific focus for their evaluations, providing assessments of specific products, services, art, performances, *etc*. Beyond that, the rest of this category consists of religious blogs/sermons (8.5 percent) and advice documents (4.5 percent). Here again, we see the rarity of a register considered to be important in corpora of published written texts: letters to the editor comprise only 0.3 percent of the opinion general category, while overt advertisements (with more than seventy-five words of prose, see Section 5) are extremely rare.

5.4 Other general registers

The other five general register categories (Interactive Discussion, How-to/Instructional, Informational Persuasion, Lyrical and Spoken) occurred much less frequently than the three major categories of Narration, Informational Description/Explanation, and Opinion. However, it is clear from Table 8 that these registers each comprise one or two especially important sub-register categories. For example, Discussion forums and Question/Answer forums are especially important, making up almost 90 percent of the Interactive Discussion category. Similar to blogs, these are specialised web registers not found in print media.

==Insert Table 8 about here==

Most documents in the Lyrical category consisted of song lyrics, while interviews were especially prevalent in the Spoken category. Not surprisingly, how-to explanations, recipes (i.e., a special category of How-to that is intended specifically for cooking), and more formal instructions for other processes dominate the How-to/Instructional category.

The Informational Persuasion register consists mostly of the sub-register 'description with intent to sell'. These are similar to infomercials, in that the primary content is information about a place or product, while the underlying motivation is to persuade the reader to visit that place or purchase the product. Although this is the dominant sub-register of Informational Persuasion, these documents are not especially prevalent on the web generally (they account for only about 1.6 percent of the entire corpus). These documents can be regarded as a kind of hybrid register, combining the communicative purposes of informing/describing/explaining with a persuasive goal. As a result, users actually had difficulty in agreeing on this register categorisation. As we show in the following section, there were other important differences in the extent to which these categories were perceptually well-defined for users.

5.5 Extent to which the registers are perceptually well-defined

The preceding sections focus on the large number of documents that users were able to agree on. An alternative perspective is to consider the perception of the registers themselves, investigating the extent to which these categories are perceptually well-defined for users. That is, if one user codes a document as an instance of a register, what is the likelihood that the other three users will perceive this same register?

Table 9 presents Fleiss' Kappa agreement coefficients for each register, showing that there are large differences in the extent to which the categories are perceptually well-defined. At one extreme, there is high agreement for Interactive Discussion and Lyrical documents: if one rater perceived a document as belonging to these categories, it is likely that other raters would agree.

==Insert Table 9 about here==

At the other extreme, we see a low rate of agreement for Informational Persuasion. In fact, we found that there were 4,506 documents that were coded by only one rater as an instance of Informational Persuasion as compared with only 216 documents where all four raters agreed on a coding of Informational Persuasion. Thus, this register category was especially nebulous for most raters (often coded instead as Informational Description/Explanation and/or Opinion).

In part, the results summarised in Table 9 indicate that these general register categories are not equally well-defined for end-users. But

these results also reflect the fact that many documents are not pure instances of a single register. So, for example, Table 4 above shows that the registers of Informational Description/Explanation and Opinion are perceptually well-defined: these are two of the three most common register categories in our corpus, and thousands of documents in our corpus were coded with complete agreement as belonging to these categories. At the same time, Table 9 shows low levels of agreement for the overall coding of these two categories (Fleiss' Kappa = 0.37 and 0.36). This apparently contradictory finding can be explained in part by positing the existence of 'hybrid' registers – documents that combine multiple communicative purposes in a single text. It turns out that the patterns of coding from our raters also support the existence of these hybrid register categories.

5.6 Hybrid registers

In Section 4, we noted that many web pages were coded with a 2–2 split. For example, two raters might have coded a given page as a Narrative, while two other raters classified the same page as Informational Description/Explanation. One interpretation of these splits is that they simply show a lack of agreement among raters, reflecting a lack of reliability in the register framework. However, the actual distribution of these pairings suggests a different interpretation.

In theory, there are twenty-eight different 2–2 categories that could be formed by combining the eight general register categories in our framework. So, for example, there are seven different 2–2 categories that could have been formed by combining Narrative with one of the other categories (Narrative-Spoken, Narrative-Interactive Discussion, Narrative-Informational Description, Narrative-Opinion, Narrative-Information presented with the intent to persuade, Narrative-How-to, Narrative-Lyrical). Similarly, there are twenty-one other pairings of general registers that are theoretically possible.

Given this potential, it is surprising that only seven combinations of general registers commonly occurred in 2–2 splits (see Table 10). This restricted set of recurring register combinations suggests an alternative explanation for the lack of agreement among raters: rather than reflecting a problem with the coding rubric, these common 2–2 combinations can be interpreted as evidence that these texts belong to 'hybrid' registers – registers that combine the communicative purposes and other situational characteristics of two or more general registers.

==Insert Table 10 about here==

Two of these hybrid combinations are especially important: Narrative + Informational Description/Explanation, and Narrative + Opinion. Taken together, those two combinations account for about 60

percent of all hybrid documents (or about 7 percent of the entire corpus). Informational Description/Explanation is also important and combines with Opinion (about 13 percent of 2–2 hybrids), Informational Persuasion (about 7.5 percent of 2–2 hybrids) and How-to/instructional documents (about 6.2 percent of 2–2 hybrids).

Table 11 shows that these same three general register categories (Narrative, Informational Description, and Opinion) dominate the 2–1–1 hybrid register categories: about 37.5 percent of all 2–1–1 hybrids were coded as combinations of these three registers, while most of the other recurrent 2–1–1 categories include two of these three general registers.

==Insert Table 11 about here==

Tables 10 and 11 show that it is especially common to combine narrative and informational purposes in the same document. Some of these documents are biographical recounts that also describe a current social situation. For example, Text Sample 1 provides a brief biographical history of David and Jackie Segal, while also describing their current lifestyle and specifically their house.

Text Sample 1:[12] Narrative + Informational Description

Imagine you are rich. Really, seriously rich. So rich that you can afford to build a 90,000 acre dream house, based on an actual French palace -- the Palace of Versailles. Then imagine losing everything.

That's what happened to David and Jackie Segal, one of America's richest couples. He made billions from his time-share business, Westgate Resorts, selling (ironically) hundreds of Americans their idea of the American Dream -- luxury lifestyles at an affordable rate.

David married former beauty queen Jackie, 31 years his junior, in 2000. Roll on seven years and millions of dollars later, and the couple are intent on re-creating the Palace of Versailles in Florida, thus building America's Biggest House.

With 10 kitchens, 30 bathrooms, two tennis courts, a baseball field, two swimming pools and an ice rink, it was to be twice as big as the Whitehouse. Not too shabby, eh?

Filmmaker Lauren Greenfield decided to make a documentary about the billionaire couple's project. She was given incredible access to the couple, their eight children, their 26,000 acre Florida mansion home and, of course, the beginnings of their mammoth project.

'Why am I building the biggest house in America? Because I can,' a smug David told the camera.

[12] Source: http://www.graziadaily.co.uk/conversation/archive/2012/09/04/the-queen-of-versailles--exclusive-clip.htm

But mid-way through the filming process, the recession hit, time shares went bust and they lost millions. David was forced to sack thousands of his employees and Jackie had to forgo her $1 million a year clothing allowance, swapping YSL for Walmart.

Plans to build their dream home were scuppered and who could afford to buy it? $5 million worth of marble is still boxed away in the basement.

Text Sample 2 illustrates a different type of narrative-informational hybrid document, combining a personal travel narrative with an informational description of the 'Way of the Roses' biking route in England.

Text Sample 2:[13] Narrative + Informational Description

"Let me guess," said the stationmaster at Lancaster as he showed me where to stow my bike on the connecting train to Morecambe. "You wouldn't be cycling to Bridlington, by any chance?" When I replied in the affirmative his small audience on the platform were most impressed. At his accuracy, I mean, not my pedalling power. "It's quite simple really," he explained. "Anyone taking a bike to Morecambe must be going to Bridlington. This train never saw any cyclists for donkey's years, now we get dozens of them and they are all doing the same thing."

Beyond a shared desire to turn back the holiday clock by about 70 years, not much would appear to link Morecambe and Bridlington, but now a coast-to-coast cycle route does, and the Way of the Roses is evidently becoming quite popular. Even in March I met others riding it and in under two years, cafes and hotels all along its 170-mile length have begun to sprout "cyclists welcome" signs.

The name is a slight misnomer, since all but the first 20 miles are in the White Rose county and the route only touches the outskirts of Lancaster. A short detour to the city centre would be perfectly possible, but extra mileage is never an appealing prospect for the cyclist looking to fit in 78 miles on the first day.

In order to complete the route in two days I had booked overnight accommodation at Ripon. I was aware most of the hills would come on the first day -- and all too painfully aware by the end of it -- but I felt I had to aim for something close to halfway to avoid a 100-mile ride on the second. You don't want to be going that far with bags on your bike.

The route is superbly signposted throughout, so much so that you can almost leave the map in your pocket. After following the Lune upstream for a few miles, the flat top of Ingleborough pops into view to give you a chunk of Yorkshire to aim for. Just below Ingleton, the

[13] Source: http://www.guardian.co.uk/travel/2012/may/05/british-bike-rides-long-distance

```
distinctive two-note call of the curlew accompanied me
across the county border -- the sort of perfect moment
that lingers long in the mind after the ride is over to
remind you why you do this sort of thing.
```

It is perhaps not surprising that opinionated purposes are also commonly combined with narrative or informational purposes. In particular, personal blogs commonly combine narrative and opinionated purposes. There are many particular ways in which these general purposes are combined in blogs – for example, in a news report presented from a particular perspective, an argumentative editorial illustrated with narratives, or a movie review that also recounts much of the plot (see Vásquez's, 2012, discussion of involvement and narrativity in opinionated consumer reviews of hotels). Similarly, informational/descriptive texts often incorporate opinionated discussion, such as a business report on a corporation that begins with an explicit disclaimer that the blog represents 'personal opinions', even though the text is mostly a simple report of financial information.

Three-way splits, summarised in Table 11, suggest that many documents actually combine multiple communicative purposes. The most frequent three-way hybrid is Narrative + Opinion + Description. For example, one document was coded as a News report/blog (two raters), a Description of a person (one rater) and an Opinion blog (one rater). The title of this text is enough by itself to suggest the triad of characteristics recognised by raters: *On the road*: *Bradley Wiggins and Team Sky have made Tour de France history – it's been emotional*. This text is a blog post that recounts a recent news story (Narrative), describes a team of athletes (Description), and recounts the emotions and attitudes of the author (Opinion).

A different kind of hybrid register is also extremely common on the web: pages with text followed by reader comments. Table 12 shows that this type of hybrid can occur with any of the non-interactive written registers.[14] However, it is interesting to note that reader comments are much more likely with some registers than others. In particular, Narrative, Opinion, How-to, and Informational Persuasion documents are commonly followed by reader comments (27–37 percent of the time), while comments are much less likely in response to Informational, Lyrical or Spoken documents (7–19 percent of the time). (Interactive Discussions are excluded from consideration because, by definition, they include reader comments.) Blogs are especially likely to include reader comments: 55 percent of narrative personal blogs, 49 percent of opinion blogs and 22 percent of informational blogs include comments from readers.

[14] This option is not applicable to written interactive discussions, which, by definition, incorporate reader comments. We are not sure why transcribed texts of spoken events are not followed by reader comments in our sub-corpus.

==Insert Table 12 about here==

5.7 Hybrid sub-registers

As discussed in relation to Table 3 (Section 4), raters are commonly split in their coding of sub-register categories: 37.6 percent of the documents in our corpus were coded as 2–2 or 2–1–1 combinations at the sub-register level. However, unlike the systematic nature of hybrid combinations at the general register level, these sub-register combinations are not highly patterned. Thus, Table 13 lists the seven 2–2 sub-register combinations that occurred more than 100 times in our corpus. Taken together, those recurrent combinations account for only 46 percent of all 2–2 sub-register combinations. The remaining 54 percent of these documents belong to more idiosyncratic combinations of sub-registers that were assigned less frequently by raters. Overall, there are 269 different 2–2 sub-register combinations that were assigned by raters in the data, and fifty-three of those combinations occurred ten or more times. For example, there were twenty-four different 2–2 combinations involving 'personal blogs' combined with another sub-register (e.g., news reports, travel blogs, advice, informational description, informational blog, opinion blog, reviews).

==Insert Table 13 about here==

The situation is even less systematic for 2–1–1 sub-register combinations. Table 14 lists the nineteen combinations that occurred in more than 100 documents in the corpus, but these account for only about 24 percent of all 2–1–1 sub-register combinations. Overall, there are 2,432 different 2–1–1 sub-register combinations attested in the corpus, and 275 different 2–1–1 combinations that occurred ten or more times.

==Insert Table 14 about here==

At the same time, raters agreed on the sub-register category of 51 percent of the documents in our corpus (see Table 3). A more detailed consideration shows that the sub-registers within some general categories were relatively transparent, while others were highly problematic. Raters usually had no difficulty in agreeing on the specific sub-register of documents within the general categories of Spoken, Lyrical and Interactive Discussion. In addition, raters agreed on the sub-register category of 90 percent of the documents within the general category of Informational Persuasion, and, surprisingly, they agreed on the sub-register of 84 percent of the documents classified as Narrative. At the other extreme, raters found it difficult to agree on the specific sub-register of Informational Description/Explanation documents: only 43 percent of those documents had majority agreement on a specific sub-register category.

Taken together, these findings replicate previous research, which has repeatedly documented problems in determining the specific sub-register of web documents. At the same time, these results show that many web documents can be reliably classified for sub-register categories. In ongoing research, we are exploring the possibility of grouping sets of documents coded with less frequent hybrid combinations, based on their shared situational characteristics. In this way, we hope to provide linguistic descriptions of both the general registers found on the web, as well as the most common specific sub-register categories.

6. Summary and future directions

The approach for register classification adopted here – a bottom-up hierarchical framework based on underlying situational characteristics – allows us to account for the register characteristics of most web pages. Raters generally agree on the general register category of about 69 percent of the web pages included in our corpus (see Table 3). About another 29 percent of the documents in our corpus can be regarded as 'hybrid' registers belonging to a few combinations that occur commonly on the web (e.g., Narration + Information Description; Narration + Opinion; see Tables 11 and 12). Taken together, these results indicate that 80–90 percent of web pages can be meaningfully described for their (hybrid) register characteristics.

The general register categories that we used are mostly associated with different general communicative purposes (e.g., narrating, informing and giving opinions). These are quite different in nature from the tidy register categories that are usually employed in written corpus designs (e.g., academic research articles or newspaper editorials). As described in Section 1, this reflects a fundamental difference between the discourse domains of published written texts *versus* searchable web documents. One consequence of this difference is that register distinctions are considerably more difficult to determine for web documents than for published written texts. However, it also appears that the register distinctions defined in terms of basic communicative purposes are not necessarily simple, because many texts combine multiple purposes. For this reason, it is not surprising that the web registers that emerged from our analysis include a set of 'hybrids'.

The interesting finding, though, is that only a few general registers and a few hybrid combinations dominate the documents found on the searchable web. These are not necessarily the most salient registers or the ones that most users would predict to be especially common. For example, news/sports reports/blogs are especially prevalent on the searchable web, making up about 21 percent of the total documents in our corpus. Various kinds of informational descriptions / explanations are also common (about 14 percent of the total), as well as opinionated texts (about 11 percent of the total). The prevalence of narrative, informational descriptive/explanatory,

and opinionated registers is even higher if we include hybrids that combine these communicative purposes: over 75 percent of all web documents. In contrast, Interactive Discussions and Forums, How-to/Procedural documents, Lyrical, and Spoken transcriptions are all much less frequent.

It is perhaps not surprising that our research findings show that blogs are probably the quintessential register of the searchable web, comprising 20–25 percent of our corpus. Blogs can vary widely in their situational characteristics and communicative purposes, and, as a result, specific blog sub-registers were categorised under several of our general registers. At one extreme are the personal blogs not associated with any institution; these can serve narrative, informational or opinionated purposes, with an incredible array of specific communicative purposes. At the other extreme are institutional news/sports blogs, which are in some cases virtually indistinguishable from published news reports. Taken together, blogs provide a microcosm of the incredible range of variation found on the web.

Our prediction is that these register and hybrid-register distinctions, defined in terms of basic situational characteristics, will correspond to systematic patterns of linguistic variation. In our on-going research, we are exploring the lexico-grammatical characteristics of these categories to document systematic linguistic patterns of register variation on the web. These analyses include both detailed investigations of particular linguistic features (e.g., stance devices) as well as a multi-dimensional analysis to identify the underlying linguistic parameters of variation. Building on those analyses, we plan to undertake predictive research for the purposes of automatic register (genre) identification.

The first step, though, is to develop a taxonomy of the registers found on the web. By adopting a bottom-up user-driven approach to document classification, based on analysis of a large corpus of web documents randomly sampled from the searchable web, we hope to have taken an important step towards achieving that objective.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. 1147581. We also thank Anna Gates and Rahel Oppliger for their help with the pilot testing of register classification schemes.

References

Baroni, M. and S. Bernardini. 2004. 'BootCaT: bootstrapping corpora and terms from the web' in Proceedings of LREC 2004, pp. 1313–16. Lisbon: ELDA.

Baroni, M., S. Bernardini, A. Ferraresi and E. Zanchetta. 2009. 'The WaCky wide web: a collection of very large linguistically processed web-crawled corpora', Language Resources and Evaluation 43 (3), pp. 209–26.

Biber, D. and S. Conrad. 2009. Register, Genre and Style. Cambridge: Cambridge University Press.

Crowston, S. 2010. 'Problems in the use-centered development of a taxonomy of web genres' in A. Mehler, S. Sharoff and M. Santini (eds) Genres on the Web: Computational Models and Empirical Studies, pp. 69–86. New York: Springer.

Egbert, J. and D. Biber. 2013. 'Developing a user-based method of web register classification' in S. Evert, E. Stemle and P. Rayson (eds) Proceedings of the 8th Web as Corpus Workshop (WAC-8) @Corpus Linguistics 2013, pp. 16–23.

Egbert, J., D. Biber and M. Davies. 2014. 'Developing a Bottom-up, User-based Method of Web Register Classification. Submitted to JASIST.

Fletcher, W.H. 2012. 'Corpus analysis of the World Wide Web' in C.A. Chapelle (ed.) Encyclopedia of Applied Linguistics, pp. 1339–47. Oxford: Wiley-Blackwell.

Guadagno, R.E., B.M. Okdie and C.A. Eno. 2008. 'Who blogs? Personality predictors of blogging', Computers in Human Behavior 24, pp. 1993–2004.

Herring, S.C., L.A. Scheidt, E. Wright and S. Bonus. 2005. 'Weblogs as bridging genre', Information Technology and People 18, pp. 142–71.

Kilgarriff, A. and G. Grefenstette. 2003. 'Introduction to the special issue on the web as corpus', Computational Linguistics 29, pp. 333–47.

Landis, J.R. and G.G. Koch. 1977. 'The measurement of observer agreement for categorical data', Biometrics 33, pp. 159–74.

Rehm, G., M. Santini, A. Mehler, P. Braslavski, R. Gleim, A. Stubbe, S. Symonenko, M. Tavosanis and V. Vidulin. 2008. 'Towards a reference corpus of web genres for the evaluation of genre identification systems' in N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis and D. Tapias (eds) Proceedings of the 6th Language Resources and Evaluation Conference, pp. 351–8. Marrakech, Morocco.

Rosso, M.A. and S.W. Haas. 2010. 'Identification of web genres by user warrant' in A. Mehler, S. Sharoff and M. Santini (eds) Genres on the Web: Computational Models and Empirical Studies, pp. 47–68. New York: Springer.

Santini, M. 2007. ‘Characterizing genres of web pages: genre hybridism and individualization’ in R.H. Sprague (ed.) Proceedings of the 40th Hawai‘i International Conference on System Sciences (HICSS-40), pp. 1–10. Hawai‘i.

Santini, M. 2008. ‘Zero, single, or multi? Genre of web pages through the users’ perspective’, Information Processing and Management 44, pp. 702–37.

Santini, M. and S. Sharoff. 2009. ‘Web genre benchmark under construction’, Journal for Language Technology and Computational Linguistics 25 (1), pp. 125–41.

Sharoff, S. 2005. ‘Creating general-purpose corpora using automated search engine queries’ in M. Baroni and S. Bernardini (eds) WaCky! Working papers on the Web as Corpus, pp. 63–98. Gedit, Bologna.

Sharoff, S. 2006. ‘Open-source corpora: using the net to fish for linguistic data’, International Journal of Corpus Linguistics 11 (4), pp. 435–62.

Sharoff, S., Z.K. Wu and K. Markert. 2010. ‘The Web library of Babel: evaluating genre collections’ in N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner and D. Tapias (eds) Proceedings of the Seventh Language Resources and Evaluation Conference (LREC 2010), pp. 3063–70. Malta.

Sindoni, M.G. 2013. Spoken and Written Discourse in Online Interactions: A Multimodal Approach. London: Routledge.

Vásquez, C. 2012. ‘Narrativity and involvement in online consumer reviews: the case of *TripAdvisor*’, Narrative Inquiry 22 (1), pp. 105–21.

Vidulin, V., M. Luštrek and M. Gams. 2009. ‘Multi-label approaches to web genre identification’, Journal for Language Technology and Computational Linguistics 24 (1), pp. 97–114.

Appendix A: sample e-mail messages from MTurk raters

E-mail messages from MTurk raters commonly sought clarification in the analysis of documents. This reflected raters' interest in the task and their concern with the quality of their work. Here is a sample of typical messages that we received (names have been changed to pseudonyms):

> Hi,
> I marked this one as "did not load" because viewing the content requires registration, otherwise it redirects to the site's main page. If I categorized it incorrectly, please let me know so I can get it right in the future.
> Thanks, Sam

> Hey Jesse, I have a question about a HIT I just completed for you. I had to classify this webpage http://backdoorbroadcasting.net/2009/09/pamela-sorensen-you-had-to-be-there-finding-meltzer-on-the-page/
> and I have one or two questions about it. The text on the page is mostly a background on an author and her credentials. Then the rest of the information on the page is a formal speech and then a question/answer session between the author and interviewers. I was unsure of how to classify it but went with "originally spoken" and "formal speech."
> Is this correct or would you rather have it in another category?
> Thanks. Rita

> Hi Jesse,
> Quick clarification on the hit I just did at:
> http://blog.transformerdesign.co.nz/aussies-hired-to-sell-nz-tourism/
> It has a twitter feed embeded on the actual page. I'm assuming that this would not be considered "reader comments at the end"
> I checked neither in this case. If embeded twitter feeds should be considered comments, please advise.
> Thanks again, John

> Good evening,
> This particular URL: https://accounts.google.com/ServiceLogin?service=blogger&hl=en&passive=86400&continue=http://www.blogger.com/blogin.g?blogspotURL%3Dhttp://www.anhaidao.com/2011/09/down-for-weekend.html%26zx%3Dz85g8fzpr3wh&ltmpl=private
> linked to a private blog that I was not able to access, so I did use the webpage not found for it as the intended page wouldn't load. I can see

that I could have gone two ways with it and I hope I didn't pick the wrong one..
Thanks again for your time. Tammy

I've been doing your hits off and on today and I just ran into one that really threw me.
http://property.mitula.co.uk/property/what-size-door-living-room
The entries are listed by multiple people, they have an intent to sell and its not really a discussion. I thought the multiple people part might be more important so I went with that. If you would let me know what you think similar things should be marked as I would appreciate it.
Thank you for your time. Edgar.

Hello,
I've been working on these HITS and while I enjoy them, I just wanted to check in and see how I am doing. I'm not quite sure how to classify Amazon sale pages (I've been classifying them as description with intent to sell w/reader's comments). Also unsure of IMDB pages. I *think* that I've got a handle on just about everything else, but don't want to do too many more without checking in.
Thanks! Lisa

Hello. I have done plenty of your hits. If you have the time to, can you let me know if I am doing fine on them? Thank you - Vicki

Hello Jesse,
I just saw the updated note on the drop down menu. I apologize for not noticing the updated text, but I have been thoroughly making sure I put things in the appropriate category.
Thank you for your time and your HITS. - Abe

Hi!
First off, thanks for the great work.
One question I came across while trying some of the web page categorization out - one question mentioned was reader comments at the end. Should we mark this for even just one reader comment, or should there be several reader comments in the available space? Paul

Table 1: Hierarchical framework based on situational parameters, used to code register characteristics of web documents

Mode	Originally written							Originally spoken
Participants	Single author or co-authors						Multiple participants	
Purpose	To narrate events	To describe information	To express opinion	To use facts to persuade	To explain instructions	To express lyrically		
Register	Narrative	Info. Description/ Explanation	Opinion	Info. Persuasion	How-to/ Instruct.	Lyrical	Interactive Discussion	Spoken
Sub-registers	-News report -Sports report -Personal blog -Historical article -Travel blog -Short story -Novel -Biography -Mag. article -Obituary -Memoir	-Describe a thing -Info. blog -Describe a person -Research article -Abstract -FAQ (info) -Legal terms -Course materials -Encyclopedia article -Tech. report	-Opinion blog -Review -Religious blog -Advice -Letter to editor -Self-help -Advert.	-Description with intent to sell -Persuasive article -Editorial	-How-to -Recipe -Instruction -FAQ (HT) -Technical support	-Lyrics -Poem -Prayer	-Discussion forum -QA forum -Responses	-Interview -Transcript -Speech -Script
Reader comments?								
Spoken quotes?								

Table 2: Agreement results for the general register classification of 48,571 web documents

4 agree	3 agree	2–2 split	2–1–1 split	No agreement
17,935	15,684	5,682	8,515	755
36.9%	32.3%	11.7%	17.5%	1.6%

Table 3: Agreement results for the specific sub-register classification 48,571 web documents

4 agree	3 agree	2–2 split	2–1–1 split	No agreement
11,769	13,220	3,526	14,576	5,480
24.2%	27.2%	7.3%	30.0%	11.3%

Table 4: Frequency information for general register categories

General Register	No. of documents	Percent
Narrative	15,171	31.2
Informational Description/Explanation	7,042	14.5
Opinion	5,452	11.2
Interactive Discussion	3,104	6.4
How-to/Instructional	1,126	2.3
Informational Persuasion	794	1.6
Lyrical	605	1.2
Spoken	325	0.7
Hybrid (see below)	14,197	29.2
No agreement	755	1.6
Total	48,571	100

Table 5: Frequency information for narrative sub-register categories

Register: Narrative	No.	Percent
News report/blog	7,967	52.5
Sports report/blog	2,444	16.1
Personal narrative blog	1,718	11.3
Historical article	206	1.4
Travel blog	128	0.8
Short story	117	0.8
Novel	32	0.2
Biographical story/history	33	0.2
Magazine article	18	0.1
Obituary	5	0.03
Memoir	1	0
Other	0	0
No majority agreement on sub-register	2,502	16.5
Total	15,171	100

Table 6: Frequency information for informational sub-register categories

Register: Informational Description/Explanation	No.	Percent
Description	1,584	22.5
Encyclopedia article	465	6.6
Informational blog	337	4.8
Description of a person	236	3.4
Research article	197	2.8
Abstract	147	2.1
FAQ about information	108	1.5
Legal terms and conditions	103	1.5
Course materials	44	0.6
Technical report	6	0.1
Other	18	0.3
No majority agreement on sub-register	3,797	53.9
Total	7,042	100

Table 7: Frequency information for opinion sub-register categories

Register: Opinion	No.	Percent
Opinion blog	2,064	37.9
Review	1,145	21.0
Religious blog/sermon	461	8.5
Advice	246	4.5
Letter to the editor	18	0.3
Self-help	3	0.06
Advertisement	2	0.04
No majority agreement on sub-register	1,513	27.8
Total	5,452	100

Table 8: Frequency information for other sub-register categories

Register	No.	percent
Interactive Discussion		
Discussion forum	1,810	58.3
Question/answer forum	911	29.3
Reader/viewer responses	7	0.2
Other	2	0.06
No majority agreement on sub-register	374	12.0
Total	3,104	100
How-to/Instructional		
How-to	544	48.3
Recipe	126	11.2
Instructions	70	6.2
FAQ	17	1.5
Technical support	9	0.8
Other	0	0
No majority agreement on sub-register	360	32.0
Total	1,126	100
Informational Persuasion		
Description with intent to sell	691	87.0
Persuasive article or essay	14	1.8
Editorial	8	1.0
No majority agreement on sub-register	81	10.2
Total	794	100
Lyrical		
Song lyrics	527	87.1
Poem	54	8.9
Other	4	0.7
No majority agreement on sub-register	20	3.3
Total	605	100
Spoken		
Interview	250	76.9
Transcript of video/audio	28	8.6
Formal speech	22	6.8
TV/movie script	12	3.7
Other	5	1.5
No majority agreement on sub-register	8	2.5
Total	325	100

Table 9: Fleiss' Kappa coefficients indicating the extent to which raters agreed in their perceptions of each register category

Register category	Fleiss' Kappa
Narrative	0.51
Informational Description/Explanation	0.37
Opinion	0.36
Interactive Discussion	0.86
How-to/Instructional	0.47
Informational Persuasion	0.26
Lyrical	0.82
Spoken	0.46

Table 10: General register 2+2 hybrid combinations (occurring more than 100 times in the corpus)

Two-way hybrid	Freq.	Percent of two-way hybrids
Narrative + Informational Description/Explanation	1,786	31.4
Narrative + Opinion	1,623	28.6
Informational Description/Explanation + Opinion	715	12.6
Informational Description/Explanation+ Informational Persuasion	427	7.5
Informational Description/Explanation + How-to/Instructional	351	6.2
Opinion + How-to/Instructional	157	2.8
Opinion + Informational Persuasion	153	2.7
Narrative + Other	225	4.0
Opinion + Other	113	2.0
Informational Description/Explanation + Other	91	1.6
All other 2–2 coding splits	41	0.7
Total	5,682	100.0

Table 11: General register 2+1+1 hybrid combinations (occurring more than 100 times in the corpus)

Three-way hybrid	Freq.	Percent of three-way hybrids
Narrative + Informational Description/Explanation + Opinion	3,192	37.5
Informational Description/Explanation + Opinion + Informational Persuasion	984	11.6
Narrative + Opinion + Informational Persuasion	934	11.0
Narrative + Info. Description/Explanation + Info. Persuasion	751	8.8
Informational Description/Explanation + Opinion + How-to/Instructional	607	7.1
Narrative + Informational Description/Explanation + Spoken	212	2.5
Narrative + Informational Description/Explanation + How-to/Instructional	210	2.5
Narrative + Opinion + How-to/Instructional	196	2.3
Narrative + Opinion + Discussion	155	1.8
Info. Description/Explanation + How-to/Instructional + Info. Persuasion	144	1.7
Informational Description/Explanation + Opinion + Discussion	138	1.6
Narrative + Opinion + Spoken	116	1.4
All other 2–1–1 coding splits	876	10.3
Total	8,515	100.0

Table 12: Frequency information for texts containing reader comments, by register (excluding interactive discussions)

	Total	No. with reader comments	Percent with reader comments
Narrative	15,171	5,055	33.3
Informational Description/Explanation	7,042	510	7.2
Opinion	5,452	2,034	37.3
How-to/Instructional	1,126	307	27.3
Informational Persuasion	794	236	29.7
Lyrical	605	74	12.2
Spoken	325	61	18.8

Table 13: Specific sub-register 2+2 hybrid combinations (occurring more than 100 times in the corpus)

2-way Hybrid	Freq.	Percent of 2-way Hybrids
News report/blog + Letter to the editor	411	11.7
Personal blog + Advice	295	8.4
Description of a thing + Reader/viewer responses	266	7.5
News report/blog + Description of a person	217	6.2
News report/blog + Personal blog	185	5.2
Discussion forum + Other forum	145	4.1
Description of a thing + FAQ about information	103	2.9
All other 2–2 coding splits	1,904	54.0
Total	3,526	100.0

Table 14: Specific sub-register 2+1+1 hybrid combinations (occurring more than 100 times in the corpus)

3-way Hybrid	Freq.	Percent of 3-way hybrids
News report/blog + Informational blog + Advice	233	1.6
News report/blog + Opinion blog + Description with intent to sell	222	1.5
News report/blog + Description of a thing + FAQ about information	214	1.5
Description of a thing + FAQ about information + News report/blog	188	1.3
News report/blog + Opinion blog + Other Informational persuasion	157	1.1
News report/blog + Description of a thing + Advice	152	1.0
Personal blog + Informational blog + Advice	137	0.9
News report/blog + Description of a thing + Opinion blog	136	0.9
News report/blog + Description of a thing + Religious blog/sermon	132	0.9
Description of a thing + Informational blog + Legal terms and conditions	132	0.9
News report/blog + Other narrative + Opinion blog	127	0.9
Description of a thing + Informational blog + Other How-to/Informational	108	0.7
All other 2–1–1 coding splits	12,638	86.7
Total	14,576	100.0