

Toppan Best-set Premedia Limited	
Journal Code: ASI	Proofreader: Mony
Article No: ASI23308	Delivery date: 15 May 2014
Page Extent: 15	

# Developing a Bottom-up, User-Based Method of Web Register Classification

**Jesse Egbert and Douglas Biber**

English Department, Northern Arizona University, Box 6032, Flagstaff, AZ 86011. E-mail: {Jesse.Egbert, Douglas.Biber}@nau.edu

**Mark Davies**

Department of Linguistics and English Language, Brigham Young University, 4071 JFSB, Provo, UT 84602. E-mail: Mark\_Davies@byu.edu

This paper introduces a project to develop a reliable, cost-effective method for classifying internet texts into register categories, and apply that approach to the analysis of a large corpus of web documents. To date, the project has proceeded in two key phases. First, we developed a bottom-up method for web register classification, asking end users of the web to utilize a decision-tree survey to code relevant situational characteristics of web documents, resulting in a bottom-up identification of register and subregister categories. We present details regarding the development and testing of this method through a series of 10 pilot studies. Then, in the second phase of our project we applied this procedure to a corpus of 53,000 web documents. An analysis of the results demonstrates the effectiveness of these methods for web register classification and provides a preliminary description of the types and distribution of registers on the web.

## Introduction

The World Wide Web is a tremendous resource of information that is growing at an accelerated rate. The identification of register (or genre) is particularly important for natural language processing (NLP) applications in computational linguistics, improving the performance of word disambiguation software, taggers, parsers, and information retrieval tools. Linguists have also recently begun to use the web as a corpus for studies of linguistic variation and use. However, the unique nature of the different types of language used on the web remains unclear. Without a clear understanding of the linguistic variability of internet texts we are severely limited in our ability to use this powerful resource for linguistic and NLP research.

Received January 8, 2014; revised March 12, 2014; accepted April 1, 2014

© 2014 ASIS&T • Published online in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/asi.23308

In order to better understand the language of the internet, it needs to be systematically classified into registers.

This article introduces a bottom-up, user-based method of classifying web documents into register/genre categories. After a brief introduction to the concepts of *register* and *genre*, we survey previous attempts to identify the register/genre of web documents, including automatic and user-based approaches. We then introduce our project to develop a user-based method of web register classification, describing the two major phases of the project completed to date.<sup>1</sup> In the first phase, we developed a comprehensive framework for the classification of register and subregisters, and we developed a methodological approach and web tool that permits user-based classification of internet texts. Then, in the second phase, we recruited 908 end users to code a corpus of 53,000 web documents for a range of register characteristics, resulting in the identification of register and subregister categories.

## Literature Review

### *Registers and Genres*

Over the past three decades, register has emerged as one of the most important predictors of linguistic variation, and a wide range of registers have been described and compared (see the surveys of previous research in Atkinson & Biber, 1994; Biber & Conrad, 2009, pp. 271–95). The terms *register* and *genre* both have long and varied histories of use in this tradition. Some researchers distinguish between the two terms, and there have been attempts to tease apart their use in previous literature (see, e.g., Biber, 1995, pp. 7–10; Lee,

<sup>1</sup>It is important to note here that, whereas some previous studies may employ the term “user” to mean user of a specific genre, we define “user” more generally to mean end user of the internet.

1  
2  
3  
4  
5  
6

7  
8  
9  
10

2001). However, for much of the existing research on text type variation this distinction is not relevant, and the choice between genre and register comes down to personal preference or tradition. For example, early research in the tradition of multidimensional analysis was framed in terms of genre variation (e.g., Biber, 1988), while research since the early 1990s has used the cover term register (e.g., Biber & Finegan, 1994; Biber, 1995).

Biber and Conrad (2009) develop a framework for distinguishing between registers and genres (see, e.g., Chapter 1, particularly Section 1.4). Registers are text varieties that are initially defined based on their situational characteristics (e.g., participants, interactivity, communicative purposes, topic), and they can then be analyzed in terms of their pervasive lexico-grammatical linguistic characteristics. Those linguistic analyses have a functional basis. That is, a linguistic feature is commonly found in a register, because it is required functionally for that situational context of use. (For example, first and second pronouns are required in interactive registers.) In contrast, genres are text varieties that are defined based on their conventional structures. For example, business letters conventionally begin with a salutation (Dear xx) and end with a politeness expression (e.g., Sincerely). Newspaper articles conventionally begin with a concise title and identification of the place where the story occurred.

Our present project on web documents adopts a register perspective, building on the tradition of research carried out over the past two decades that initially defines text categories based on their situational characteristics, and then analyzes the linguistic characteristics of those categories based on the quantitative distributions of pervasive lexico-grammatical features. For that reason, we use the cover term “register” for our project. However, there has been a separate research tradition on Automatic Genre Identification (AGI), and we retain the term genre for our discussion of work in that area.

### *Automatic Genre Identification*

There have been several previous attempts to automatically identify the register/genre categories of internet texts, carried out under the rubric of AGI. Several AGI studies have achieved high accuracy rates (e.g., Lindeman & Littig, 2010; Santini, 2010; Sharoff, Wu, & Markert, 2010). However, there are concerns about the methods and corpora used for these studies, raising questions about the generalizability of the findings.

One important issue concerns the test corpus used to evaluate automatic classification efforts, where the register/genre category of each document has been manually classified before assessing whether the automatic methods can identify those same categories. The dominant methodological approach used to develop these test corpora relies on ratings from expert coders, often the researcher carrying out the project (see Sharoff et al., 2010). This approach has been justified based on the belief that evaluations of reliability are

not necessary if an expert in genre-related research codes all texts in the corpus. However, the few cases where interrater reliability has been measured show that it tends to be quite low, even among linguists. This is especially true for corpora comprised of randomly extracted web texts (Sharoff et al., 2010, p. 6).

In addition, the nature of the corpora used for tests of AGI raises questions about the potential accuracy of AGI models applied to a larger random sample of internet texts. For example, AGI researchers seldom know whether the sample in a given corpus represents the full population of internet texts or whether the texts within a given genre class represent the variability of the descriptors included in the model (see Santini & Sharoff, 2009).

A final limitation of expert genre classification is that the distinctions made by experts do not necessarily represent genre/register categories that are meaningful to end users. To address this concern, some studies adopt an alternative approach to the manual coding of web documents, relying on actual internet users rather than “experts.” However, given the aforementioned problems that “experts” have identifying web genre/register categories, it is not surprising that nonexpert web users also vary in their understanding of genre/register labels (see Crowston, Kwasnik, & Rubleske, 2010), and previous research has shown that reliability among end users is often unacceptably low (Rosso & Haas, 2010). There are many possible explanations for the low agreement among users. These include the meaningfulness of the categories, the level of register specificity, the multifaceted nature of registers, “fuzziness” in register taxonomies, and the existence of hybrid texts. Each of these issues will be discussed briefly in the next section.

### *Issues in User-Based Register Classification*

Rosso (2008, p. 1057) has argued that in order for a genre/register category to be useful it must be meaningful to users who are knowledgeable in the use of that text type. Many attempts have been made to develop user-based register taxonomies based on user input, often in the form of sorting tasks in which users are given many web documents and asked to sort them into groups with similar documents. In these tasks the users are also often asked to assign labels to each group. However, many researchers have found it challenging to find agreement among users in the labels they assign to these text categories (see, e.g., Rosso & Haas, 2010; Crowston et al., 2010).

Another related issue is that of register abstraction or specificity. Researchers have found, unsurprisingly, that interrater agreement tends to be lower for more specific register categories (e.g., Haas & Grams, 1998). Other research suggests that interrater disagreement is often hierarchical in nature. In other words, when two raters disagree at a low level of register abstraction, researchers have noticed that the two choices are often subregisters of a single, more general register (e.g., Crowston & Williams,

2000, p. 205; Roussinov et al., 2001, p. 5). However, to date the research on register hierarchies has been qualitative and exploratory, and much more research is needed in this area.

One proposed approach to this issue is to develop a register classification framework based on facets rather than register labels (Crowston & Kwasnik, 2004; Rosso, 2008). According to Crowston and Kwasnik (2004), a facet is essentially a parameter (e.g., form, content, source, style, implied use) that can be used to describe a register from a particular perspective (see pp. 4–5). However, others have questioned the feasibility of this approach, and no large-scale attempt at “faceted classification” has been previously attempted with internet texts (Kwasnik, Crowston, Chun, D’Ignazio, & Rubleske, 2006; Rosso, 2008).

Although interest in a faceted approach to web register classification is a relatively recent development, linguists have been analyzing registers using a faceted approach for more than 25 years. For example, Biber (1988) developed a framework to distinguish among registers based on a comprehensive set of relevant situational parameters. These parameters can be thought of as situational facets that work together to comprise the core definition of a register (see Biber & Conrad, 2009, Chap. 2). This highly effective approach has been used to classify texts in hundreds of register studies (see the survey in Biber & Conrad, 2009, pp. 271–295).

Finally, register classification can be challenging because of the existence of fuzzy register boundaries and hybrid text categories (see Rosso, 2008, pp. 1062–1063). Fuzzy register boundaries exist between registers that are similar in many ways and differ with regard to only a few characteristics. For example, Rosso (2008) asked users to distinguish between “personal website” and “welcome/homepage.” However, disagreement among raters seemed to stem from the lack of clear distinguishing characteristics between those two registers (p. 1067).

In contrast to fuzzy registers, register hybrids occur when one webpage has the characteristics of more than one register (see, e.g., Santini, 2007, 2008; Vidulin, Luštrek, & Gams, 2009). For example, Rosso (2008) found disagreement among users looking at online descriptions of a book; users disagreed on whether to label the webpage as an “article” or as a description of a “product for sale/shopping,” because the document contained a description of the book, excerpts from the book, review comments about the book, and also a link to purchase the book. Although several studies have documented the existence of fuzzy registers and register hybrids, researchers have not yet found effective methods of dealing with them in practice.

In summary, while there has been widespread interest in identifying the genre/register category of web documents, there are also still numerous methodological challenges that need to be resolved. In the following sections we describe our efforts to address these issues and develop a more comprehensive and reliable analysis of the registers found on the web.

## Study Aims

The ultimate goal of this project is to provide a comprehensive description of the linguistic patterns of register variation on the web. To achieve that goal, we developed and applied a user-based method for web register classification, and we report on those two phases of the project in the present paper. It is worth noting that our goal differs from the aims of many of the studies cited above: our primary focus is describing linguistic variation on the web, whereas the primary aim of most previous studies is improving NLP applications of web data, specifically web search.

To address the methodological challenges encountered in previous research, we developed a hierarchical register framework and a decision-tree survey that takes into account a range of relevant situational parameters (or facets). The multipage survey tool allows users to report information about relevant situational facets, one at a time, before choosing a subregister category from a list of options. In this way, we gain the information necessary to classify texts hierarchically into both general register categories and specific subregister categories. In addition, the multifaceted structure of the register framework also makes it possible to request situational information from users in a manageable and incremental manner.

After numerous rounds of pilot testing and revision, we applied our methodological approach to a large corpus of 53,000 web documents. Each document was coded by four end users of the internet, allowing us to assess the reliability of the framework as well as the extent to which particular registers are well defined. (Note that we employ the term “user” to refer to users of the internet, rather than users of a specific register/genre, as in some previous research (e.g., Crowston et al., 2010).) In addition, coding with multiple end users allowed us to (a) determine the fuzziness of register and subregister categories and (b) identify texts that are hybrids between two or three register or subregister categories.

In sum, the user-based method of web register classification introduced here makes it possible to (a) classify texts into register categories, (b) account for a wide range of situational parameters (facets), (c) classify texts hierarchically at different levels of specificity, (d) measure the fuzziness of register categories, and (e) identify and quantify hybrid register categories.

## Phase 1: Developing a Bottom-up Method for Web Register Classification

This section describes the development of our bottom-up approach for the classification of internet texts into register categories. Beginning with an analytical framework developed over the last 20 years (see, e.g. Biber, 1995; Biber et al., 1999; Biber & Conrad 2009), we define “register” as a language variety that is “associated with a particular situation of use including particular communicative purposes” (Biber & Conrad, 2009, p. 6). Whereas previous approaches

to web register classification have usually disregarded the difference between linguistic and nonlinguistic factors in their definitions of registers/genres, this study first establishes register categories according to the situation in which language is used. Subsequent analyses are then planned to investigate the linguistic characteristics of these registers (see our discussion in the concluding section.)

### Corpus

The corpus used for the study was extracted from the Corpus of Global Web-based English (GloWbE) (see <http://corpus2.byu.edu/glowbe/>). The GloWbE corpus contains ~1.9 billion words in 1.8 million web documents, collected by using the results of Google searches of highly frequent English 3-grams (e.g., *is not the, and from the*). Although nearly half of the GloWbE corpus was sampled from Google Blogs, the sample in this study was drawn from the “General” portion of GloWbE. Many previous web-as-corpus studies have used n-grams as search engine seeds (see, e.g., Baroni & Bernardini, 2004; Baroni, Bernardini, Ferraresi, & Zanchetta, 2009; Sharoff, 2005, 2006). After identifying webpages from the search results, we downloaded them using HTTrack (<http://www.httrack.com>). We then removed all nontextual material (HTML and boilerplate) from the webpages using JusText (<http://code.google.com/p/justext>) in order to prepare the corpus for future linguistic analysis.

For the present project we randomly extracted 53,000 documents from the GloWbE Corpus. This sample, comprising webpages from five geographic regions (United States, Great Britain, Canada, Australia, and New Zealand), represents a large, random sample of web documents found on the searchable web.

### Developing a Register Framework for Web Document Classification

We began work on this project by attempting to develop a comprehensive taxonomy of web registers. That is, when we started to work on this project we assumed that end users would be able to directly identify the register category of web texts. Thus, our first task was to develop an initial framework that itemized the possible register distinctions found on the web. To accomplish this we began with the 78 text categories that resulted from a wiki-based collaboration among web-as-corpus experts (<http://www.webgenrewiki.org/>; see the discussion in Rehm et al., 2008). We then surveyed a random sample of 200 webpages for the purpose of capturing register categories that were new or otherwise unrepresented on that list. Based on our own previous experience with register analysis, we grouped the resulting categories into eight general register categories, with 68 subregister categories grouped under these top-level categories (see Table 1).

We then undertook a series of pilot studies to assess the extent to which end users could reliably assign internet texts

TABLE 1. Initial register framework with example subregister categories.

General register	Example subregisters
Simple Description	description of an organization, about page, course description
Technical Informational Writing	research article, technical report, abstract
Non-fiction Narrative	newspaper report, historical article, biography
Fiction/Personal Narrative	personal blog, diary, novel
Opinion/Persuasion	opinion blog, editorial, review
How-to/Procedural	frequently asked questions, self-help, recipe
Discussion	forum, chat, guestbook
Speech	interview, debate, TV/movie script

TABLE 2. Overview of the samples analyzed in the 10 pilot studies.

Study	URLs	Raters	Instrument
1	25	2	Rubric
2	25	2	Flowchart
3	25	2	Decision-tree survey
4	25	2	
5	25	3	Decision-tree survey
6	25	3	
7	25	3	Decision-tree survey
8	50	5	
9	100	4	Decision-tree survey
10	1,000	4	

to these register categories, modifying the categories themselves as needed to achieve higher rates of agreement. However, we reconsidered our basic methodological approach after achieving low interrater agreement in the first pilot study. We determined that end users could not directly identify registers and subregisters in a reliable manner, and so we instead asked coders to begin by identifying basic situational characteristics of texts (related to mode, interactivity, and communicative purpose), which would then lead coders to reduced sets of specific register categories.

Over the course of the evaluation and revision process we undertook 10 different pilot studies. Table 2 contains a brief summary of the samples analyzed in these studies.

We evaluated the results after each of these pilot studies and revised the methodological approach to address problems with the coding. Table 3 documents the major changes to the register framework and coding methods over the course of the pilot studies, along with the factors that motivated each revision.

As documented in Table 3, there were extensive changes made to the register framework during the course of the 10 pilot studies. For example, we revised our treatment of texts with extensive quoted speech and extensive reader comments, recognizing that these could be characteristics associated with any register rather than distinct registers in themselves. As a result, we revised the coding framework so that raters could note the occurrence of extensive quotes

TABLE 3. Major changes made to the initial register framework and the reasons for the changes.

Change	Reason
1. Texts composed of more than 50% spoken quotes classified as Speech	1. Users were unsure of the definition of the Speech category
2. Texts composed of more than 50% reader comments classified as Discussion	2. Reader comments are common and users did not know how to classify them
3. The rubric of register categories was modified into a flowchart based on the following binary situational variables: (1) mode (spoken/written), (2) interactivity (multi-participant/single author), (3) purpose (narrative/descriptive), and (4) factuality (opinion/objective)	3. Users were overwhelmed by the number of register categories and the many situational facets that distinguished them
4. Divided Discussion into Technical Discussion and Non-technical Discussion	4. Discussion forums were quite different depending on the intended audience: experts or nonexperts
5. Texts with reader comments are noted but not classified as Interactive Discussion	5. User agreement was low for texts with reader comments; fundamental differences were noticed between interactive discussions and texts containing reader comments
6. Texts with spoken quotes are noted but not classified as Speech	6. Many texts, especially news and sports articles, were being classified as Speech
7. Modified the flowchart into a decision-tree survey	7. Users found it difficult to work through the flowchart when all of the possibilities were presented to them simultaneously
8. Added a register category for Lyrical texts that included song lyrics, poems, and prayers	8. Users struggled to classify texts of this nature; they did not align well situationally with other registers
9. Divided Opinion/Persuasion into 2 categories: Opinion and Informational Persuasion	9. Users struggled to classify texts that were primarily informational but also had the intent to persuade the reader
10. Merged Technical Discussion and Non-Technical Discussion into one register: Interactive Discussion	10. Users found it difficult to judge the technicality of discussions
11. Began using Mechanical Turk to recruit and compensate raters	11. This allowed us to collect data in a fast, cost-effective manner
12. Condensed number of screens in the decision-tree survey by changing some of the binary options into multiple choice questions. Specifically, we (a) merged the situational variables of communicative purpose and factuality into the same page and (b) merged the mode and interactivity distinctions into the same page	12. We wanted to increase the efficiency of the instrument
13. Added a list of subregister options to select from once the register category was established based on the situational characteristics	13. We wanted to explore whether users could classify subregisters more reliably once the register category was established
14. Added a list of frequent, easy-to-identify subregister options to a dropdown menu on the first page of the survey	14. Users mentioned that certain text types were quite frequent and easy to identify and wanted a quicker way to access them
15. Merged Simple Description and Technical Informational Writing categories into one register: Informational Description/Explanation	15. It was determined that these differed primarily in the degree of technicality, and that information could be garnered from the subregister data
16. Merged Nonfiction Narrative and Fiction/Personal Narrative categories into one register: Narrative	16. These categories differed primarily in the degree of objectivity of the narrative, but that information could be garnered from the subregister data

and/or reader comments in any web text, regardless of the register category that the text was assigned to.

We also made several revisions to the register categories in our framework during the piloting process. For example, early in the process we subdivided the Discussion category into Technical Discussion versus Nontechnical Discussion; and then later in the process we determined that users were not able to reliably make this distinction, and so we merged those two categories back into the single category of Interactive Discussion.

In Step 11, we began using Mechanical Turk (MTurk), an Amazon-based crowdsourcing site, to recruit and pay raters who were actual end users of the web. This was an important innovation, which allowed us to collect massive amounts of data in a relatively quick and cost-effective way. Previous research has investigated whether results from MTurk workers are comparable to data collected using other

methods, showing that there are no significant differences between MTurk workers and participants recruited from other populations (Paolacci, Chandler, & Ipeirotis, 2010; Suri & Watts, 2011). Especially for the coding tasks required for our project, we found MTurk to be an excellent means of recruiting participants and collecting data for the classification tasks.

In addition to the application of MTurk, the most important methodological innovation we made during the course of the 10 pilot studies occurred in Step 3, when we modified the classification rubric into a decision tree based on the full set of relevant situational characteristics. This allowed users to focus on individual situational parameters, rather than trying to directly identify a register category. At this stage we introduced four major situational parameters: mode, interactivity, communicative purpose, and factuality. Although some of these distinctions were later merged (see

1 Step 12), beginning with basic situational distinctions  
2 remained the key consideration in the register classification.  
3 This change led to the introduction of a register hierarchy  
4 in the framework, with each situational parameter identify-  
5 ing a register category at a greater degree of specificity. For  
6 example, the top level distinguished between spoken and  
7 written registers, while the second level distinguished  
8 between interactive written registers versus noninteractive  
9 written registers. In Step 13 we further added lists of specific  
10 subregisters as an additional level of specificity once they  
11 had narrowed the situational characteristics of a text down to  
12 the register level. We found that reliability was relatively  
13 high on the subregister level when users chose from a  
14 limited set of 4–12 related subregisters that were possible  
15 realizations of a high-level register category (rather than  
16 choosing directly from an extended list of all 50+ subregister  
17 categories).

21 Based on the first several rounds of pilot testing, we  
22 noticed that a few subregister options were highly frequent  
23 and relatively easy to identify. Thus, to make the survey  
24 more efficient we introduced (Step 14) a dropdown menu  
25 allowing experienced coders to directly select from a list of  
26 seven common subregisters: news report/blog, sports report,  
27 opinion blog, review (product, service, movie, etc.), simple  
28 description (e.g., of a place, product, organization, program,  
29 job), description with intention to sell, and question/answer  
30 forum. This made it possible for experienced coders to clas-  
31 sify a text without working their way through the entire  
32 survey. However, during training coders were instructed to  
33 allow the survey to guide them to the most appropriate  
34 register category if they had any doubt. This preferred prac-  
35 tice was reemphasized in a note above the dropdown menu  
36 that permitted direct selection of a common subregister (see  
37 Figure 1).

## Internet Text Survey

You will be asked a series of questions about the writing on the internet page we have given you. You should focus on the text in the main body of the web page, and ignore any writing in advertisements or links. Please select the BEST answer to each question.

\* Required

Please enter your MTurk Worker ID: \*

Enter the URL for the webpage you are classifying \*

Note: If the webpage \*automatically\* redirects to a new URL then enter the new URL rather than the old one.

Enter the ID number from the bottom of the HIT \*

The text on this webpage is...

- written by one author or co-authors
- written by multiple participants in a discussion format (NOT including reader comments following an article or essay)
- originally spoken [NOT song lyrics] (interview, formal speech, transcript of video/audio recording, scripts from TV, movies, or plays, etc.)
- mostly photos or graphics (less than 50 words of text)
- webpage not available (please only select this option after trying the URL in 2 different browsers (ex: Firefox, Internet Explorer, Google Chrome))

...OR choose from one of these common registers (IF you are already certain)

NOTE: The use of this drop-down menu will be monitored and its overuse (using it most of the time) will be investigated. As there are more than 50 register categories we encourage you to please consider all options for a given text by using the options above to allow the survey to guide you to the appropriate register category. This short list of common register categories should be used only if you are 100% certain you already know the correct register.

Continue »

18

Colour online. B&W in print

19  
20

FIG. 1. Screenshot of the first page of the Google Survey instrument. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

TABLE 4. Final register framework with associated subregisters.

Register	Subregisters
Narrative	News report/blog; Sports report; Personal/diary blog; Historical article; Travel blog; Short story; Novel; Biographical story/history; Magazine article; Obituary; Memoir; Other narrative
Opinion	Opinion blog; Review; Religious blog/sermon; Advice; Letter to the editor; Self-help; Advertisement; Other opinion
Informational Description/ Explanation	Description of a thing; Informational blog; Description of a person; Research article; Abstract; FAQ about information; Legal terms and conditions; Course materials; Encyclopedia article; Technical report; Other informational description/explanation
Interactive Discussion	Discussion forum; Question/answer forum; Reader/viewer responses; Other interactive discussion
How-to/ Instructional	How-to; Recipe; Instructions; FAQ about how-to; Technical support; Other how-to/instructional
Informational Persuasion	Description with intent to sell; Persuasive article or essay; Editorial; Other informational persuasion
Lyrical	Song lyrics; Poem; Prayer; Other lyrical
Spoken	Interview; Transcript of video/audio; Formal speech; TV/movie script; Other spoken

The final register framework we established (based on the rounds of piloting summarized in Table 3) is presented in Table 4. This revised framework can be compared with the initial framework in Table 1.

#### *Developing the Instrument for User-Based Register Classification*

As mentioned above, the first instrument that we developed for this project was a simple list of register categories, asking coders to choose directly from that list. After the first round of piloting we determined that users were not able to reliably select a single register category directly from a list. Therefore, we modified the instrument to become a visual flowchart that guided the rater through a series of binary situational decisions until they had arrived at the most appropriate register category for the web document.

When we began to implement the coding procedure with actual end users, we transformed the flowchart into an online decision-tree survey. Figure 1 shows a screenshot from the first page of the online instrument, while Table 5 presents a schematic representation of the hierarchical decisions in the instrument. Coders using the instrument begin with the major situational distinctions made in the final register framework. Based on their previous choices, the instrument guides raters through a series of 2–5 pages, until their responses establish the most appropriate register category. Coders are then asked to select a subregister from a short list of options and, finally, they provide additional information about the existence of reader comments or quoted material in the document.

#### *Classification of Hybrid Texts*

In our later pilot studies, we began using multiple coders (see Table 2), and as a result we realized that we could develop a bottom-up approach to identify hybrid register categories. In other words, user classifications of web documents that achieved low reliability were often evenly split between two categories, and some of these “hybrid” categories emerged repeatedly across a sample of texts. The existence of hybrid web texts was something we had anticipated based on the findings of previous research (see, e.g., Santini, 2007, 2008; Vidulin et al., 2009), but there had not been previously tested methods to identify the particular hybrid categories commonly found on the web. By employing 4–5 raters in our later rounds of pilot testing, we found that we had a method to identify common hybrid categories (e.g., narrative-description), where two raters would agree on one category, and the other two raters would agree on a second category.

#### *Final Pilot Study Results*

In our final pilot study, we asked raters to code 1,000 web texts, using the decision-tree survey described above. Four raters coded each text in M-Turk. Approximately 3.6% of the documents in that sample were no longer available on the web, and about 3.3% of the documents were labeled as not having enough text to rate. Thus, we actually obtained data on the register categories of 931 web documents in that pilot study. Table 6 summarizes the extent to which raters agreed on the register categories of these web documents (see also Biber & Egbert, in review; Egbert & Biber, 2013).

It can be seen that a majority of the raters agreed on the register category for 62.7% of the texts; all four raters agreed on the category of 315 texts, while three of the four raters were in agreement for another 269 texts. An additional ~11% of these web documents was classified as 2-2 “hybrids”: combinations of two registers that occurred at least five times, with two coders agreeing on each of the two registers associated with the hybrid (e.g., description + narrative). An additional 18.6% could be regarded as 2-1-1 hybrids, defined as a recurrent three-way combination of register categories, e.g., narrative (2 ratings) + opinion (1 rating) + description (1 rating) that occurred at least five times. Of the 56 possible 2-1-1 hybrid combinations, only seven of them were classified as recurrent categories.

Taken together, these data show that our methodological approach made it possible to classify ~92% of the texts in the pilot subcorpus into a register or register-hybrid category. On the subregister level, these data show that about 61% of the web documents could be successfully categorized into a specific subregister category or a subregister hybrid; a majority of the raters was able to agree on the subregister of ~43% of the web texts, while a smaller proportion of the texts fell into a hybrid subregister category (17.5%).

On the whole, the results of this pilot study indicated that nonexpert web users can, to a large degree, use a

TABLE 5. Visual representation of the key situational distinctions made in the final register framework.


Text	Text can be rated						Cannot rate*
Mode	Single author or coauthors			Originally written			Originally spoken
Participants	To describe information		To express opinion	To use facts to persuade	To explain instructions	To express lyrically	Multiple participants
Purpose	Info. Description/Explanation	Opinion	Info. Persuasion	How-to/Instruct.	Lyrical	Spoken	
Register	—	—	—	—	—	—	
Subregisters	— News report	— Desc. thing	— Opinion blog	— Des. to sell	— How-to	— Lyrics	— Interview
	— Sports report	— Info. blog	— Review	— Pers. article	— Recipe	— Poem	— Transcript
	— Personal blog	— Desc. person	— Relig. blog	— Editorial	— Instruction	— Prayer	— Speech
	— Hist. article	— Res. article	— Advice	— FAQ (HT)	— T. support	—	— Script
	— Travel blog	— Abstract	— Letter to ed.	—	—	—	
	— Short story	— FAQ (info)	— Self-help	—	—	—	
	— Novel	— Legal terms	— Advert.	—	—	—	
	— Bio. story	— Course mat.	—	—	—	—	
	— Mag. article	— Enc. article	—	—	—	—	
	— Obituary	— Tech. report	—	—	—	—	
	— Memoir	—	—	—	—	—	
Reader comments?							
Spoken quotes?							

\*"Not enough text (mostly photos or graphics)" or "Site not found."



TABLE 6. Agreement results for register and subregister categories in pilot study 10.

General registers				
4 agree	3 agree	2-2 hybrid	3-way hybrid	No agreement
315	269	104	173	70
33.8%	28.9%	11.1%	18.6%	7.6%
Subregisters				
4 agree	3 agree	2-2 hybrid	3-way hybrid	No agreement
171	231	73	90	366
18.3%	24.8%	7.8%	9.8%	39.3%

 9 texts were not rated (36-“websited not found”; 33-“not enough text”).

decision-tree survey to classify web documents into general registers, subregisters, and (sub)register hybrids. We considered these agreement results to be sufficiently high to justify the application of our comprehensive register framework and analytical approach to a large-scale study of web registers.

## Phase 2: Applying the User-Based Classification Approach

### *Coding the Web Documents*

Given the encouraging results of our final pilot studies, we proceeded to the analysis of our full corpus of 53,000 web documents. The methods for this phase were identical to those used in the final pilot study reported above. MTurk was used to recruit and pay all raters. Before a rater was allowed to participate in the task, they were required to complete a 7-minute interactive tutorial video. They were then required to classify a practice URL with clear situational characteristics. Classification of this practice URL was checked for accuracy before raters were approved and awarded the qualification necessary to participate in the rating process. Each time the raters accepted a classification task, they were given a URL and a link to the Google Survey that required them to enter their unique MTurk Worker ID, the URL for the web document, and a unique URL identification number. The participants then proceeded through a series of 2–5 screens requesting information about the text on the web document they were to classify (see discussion above). A total of 908 raters participated in the task by rating at least one URL. Each rater was paid \$0.11 for each URL that they classified. Each URL was classified by four independent raters.

### *Data Analysis*

The Google Survey responses were recorded in spreadsheet format with columns for each of the responses. The final spreadsheet contained 212,000 responses (53,000 URLs × 4 ratings for each document). The first step in the

analysis was to remove the URLs that were classified by one or more raters as “website not available” or “mostly photos or graphics.” There were a large number ( $n = 3,713$ ) of URLs that were classified as “website not available.” This is due in part to a nearly 7-month period of time between the initial collection of the random sample of URLs and the coding of URLs in Phase 2. An additional 1,140 URLs were classified as “mostly photos or graphics.” The resulting data set, after eliminating the URLs in those two categories, contained 48,147 URLs. This comprises the data set that will be analyzed in the remainder of our project.

The following sections contain the results of a variety of measures used to assess the reliability and effectiveness of the methodological approach developed for identification of register and subregister categories.

## Results

### *Agreement on the Coding of Documents*

We used two measures to assess the effectiveness of our methods: Fleiss’ kappa and simple percent agreement. The second measure we used is Fleiss’ kappa, which is a measure of interrater agreement that suits the design of our study. Unlike related measures of interrater agreement (e.g., Cohen’s kappa), Fleiss’ kappa does not require that the same raters coded each of the documents in the data set. However, like Cohen’s kappa, Fleiss’ kappa accounts for chance agreement among raters, making it more robust than simple percent agreement. The overall Fleiss’ kappa was .47 for the eight register categories. This can be interpreted as an indication of “moderate agreement” according to some scholars (Landis & Koch, 1977). The Fleiss’ kappa results for each of the general register categories are reported below.

It should be noted here that the use of a single measure such as Fleiss’ kappa to measure rater agreement in our study has at least two limitations. First, this measure gives us no sense of how often raters achieved partial agreement versus perfect agreement or perfect disagreement. There is a big difference between a web document that is assigned a single register category by three of four raters and one that is not agreed on by even two raters. However, Fleiss’ kappa does not provide the information necessary to group texts according to the proportion of raters that agreed. Second, the use of Fleiss’ kappa is based on the assumption that there is a single “best” category for each of the texts in the corpus. However, the results of our analysis show that this is not the case with web documents; many of the texts can be appropriately classified into hybrid register categories. Therefore, we proceed with a more detailed investigation of the results using agreement measures that are better suited to the nature of our data and methods.

We used simple percent agreement to determine the extent to which raters achieved perfect agreement (four raters) versus majority agreement (three raters) on the register and subregister levels. The overall percent agreement results for the register and subregister levels are displayed in

TABLE 7. Frequency information for majority agreement categories, for register and subregister levels.

	Register level		Subregister level	
	#	% of URL total	#	% of URL total
4 rater agreement	17,511	36.37	11,345	23.56
3 rater agreement	15,684	32.57	13,220	27.46
No majority agreement	14,952	31.06	23,582	48.98
Total	48,147	100.00	48,147	100.00

TABLE 8. Agreement results for the register categories.

	4 Agree		3 Agree		Total (3 + 4)	
	#	%	#	%	#	%
Narrative	8,641	26.03	6,530	19.67	15,171	45.70
Informational Description/Explanation	2,991	9.01	3,627	10.93	6,618	19.94
Opinion	1,908	5.74	3,544	10.68	5,452	16.42
Interactive Discussion	2,660	8.01	444	1.34	3,104	9.35
How-to/Instructional	467	1.41	659	1.99	1,126	3.39
Informational Persuasion	216	0.65	578	1.74	794	2.39
Lyrical	525	1.58	80	0.24	605	1.82
Spoken	103	0.31	222	0.67	325	0.98
Overall	17,511	52.75	15,684	47.25	33,195	100.00

Table 7. As Table 7 shows, raters were able to agree (i.e., either all four raters, or three of the four raters) on the register category of 33,195 of the documents in our corpus—69.3% of all documents.

Table 8 displays more detailed information about these documents that raters generally agreed on, showing the agreement results for each individual register category as a percent of all 33,195 documents. The information in this table is organized as follows: The first two columns present frequency and percent for the web documents that achieved perfect agreement. (For example, 26.03% of the documents that coders were able to agree on were documents where all four coders agreed on the “narrative” category; i.e., 8,641/33,195.) The next two columns present the same information for the web documents that were agreed on by three of the four raters. The final two columns contain the combined total number of web documents that achieved agreement by three or four raters, along with the percentage of the 33,195 URLs with majority agreement accounted for by each category.

Table 7 further shows that raters were able to agree (i.e., either all four raters, or three of the four raters) on the subregister category of 24,565 of the documents in the corpus—51% of all documents. As we expected, interrater agreement was lower for the subregister categories (Fleiss’ kappa = .40), but still “moderate,” according to Landis and Koch (1977). Table 9 displays the agreement results for each subregister category, considered as a percent of these 24,565 documents.

Table 7 also shows that there was *not* general agreement among raters on the register category for many documents: there was no majority agreement on the general register category for 14,952 of the 48,147 documents in the corpus (31.06%), and there was no majority agreement on the specific subregister category for 23,582 of the 48,147 documents in the corpus (48.98%). To further investigate these results, we considered the possibility of hybrid register categories. We operationally defined two-way hybrid registers as categories comprising the same two general registers in at least 100 different documents (e.g., narrative + opinion). Three-way hybrid registers were operationally defined as combinations of three general registers that occurred in the coding of at least 100 different documents. While 100 occurrences of a particular hybrid category may be considered a low threshold for consideration as a valid hybrid category, Tables 10 and 11 reveal that there are actually only a handful of hybrid categories that meet this threshold out of the many possible hybrid combinations. Considering that there are 28 possible two-way combinations and 56 possible three-way combinations of register categories, it is surprising to find that a relatively small number of categories commonly occurred.

These lists reveal a number of interesting patterns, showing that hybrids tend to incorporate some general registers more than others. For example, it can be seen from Tables 10 and 11 that the Informational Description/Explanation register category appears in 12 of the 19 hybrid categories. By summing the frequencies for these 12 hybrid

TABLE 9. Agreement results for the subregister categories.

Subregister	4 Agree		3 Agree		Total (3 + 4)	
	#	% of 3 + 4 total	#	% of 3 + 4 total	#	% of 3 + 4 total
Narrative						
News report/blog	4,467	18.18	3,500	14.25	7,967	32.43
Sports report	1,409	5.74	1,035	4.21	2,444	9.95
Personal/diary blog	545	2.22	1,173	4.78	1,718	6.99
Historical article	52	0.21	154	0.63	206	0.84
Travel blog	25	0.10	103	0.42	128	0.52
Short story	40	0.16	77	0.31	117	0.48
Novel	7	0.03	25	0.10	32	0.13
Biographical story/history	5	0.02	28	0.11	33	0.13
Magazine article	2	0.01	16	0.07	18	0.07
Obituary	2	0.01	3	0.01	5	0.02
Memoir	0	0.00	1	0.00	1	0.00
Other narrative	0	0.00	0	0.00	0	0.00
Opinion						
Opinion blog	503	2.05	1,561	6.35	2,064	8.40
Review	554	2.26	591	2.41	1,145	4.66
Religious blog/sermon	161	0.66	300	1.22	461	1.88
Advice	32	0.13	214	0.87	246	1.00
Letter to the editor	5	0.02	13	0.05	18	0.07
Self-help	1	0.00	2	0.01	3	0.01
Advertisement	0	0.00	2	0.01	2	0.01
Informational Description/Explanation						
Description of a thing	401	1.63	1,183	4.82	1,584	6.45
Informational blog	26	0.11	311	1.27	337	1.37
Description of a person	73	0.30	163	0.66	236	0.96
Research article	47	0.19	150	0.61	197	0.80
Abstract	33	0.13	114	0.46	147	0.60
FAQ about information	29	0.12	79	0.32	108	0.44
Legal terms and conditions	46	0.19	57	0.23	103	0.42
Course materials	6	0.02	38	0.15	44	0.18
Encyclopedia article	6	0.02	35	0.14	41	0.17
Other Info. Desc./Exp.	1	0.00	17	0.07	18	0.07
Technical report	0	0.00	6	0.02	6	0.02
Interactive Discussion						
Discussion forum	1,160	4.72	650	2.65	1,810	7.37
Question/answer forum	659	2.68	252	1.03	911	3.71
Reader/viewer responses	1	0.00	6	0.02	7	0.03
Other forum	0	0.00	2	0.01	2	0.01
How-to/Instructional						
How-to	183	0.74	361	1.47	544	2.21
Recipe	59	0.24	67	0.27	126	0.51
Instructions	7	0.03	63	0.26	70	0.28
FAQ about how-to	0	0.00	17	0.07	17	0.07
Technical support	2	0.01	7	0.03	9	0.04
Other How-to	0	0.00	0	0.00	0	0.00
Informational Persuasion						
Description w/ intent to sell	200	0.81	491	2.00	691	2.81
Persuasive article or essay	0	0.00	14	0.06	14	0.06
Editorial	1	0.00	7	0.03	8	0.03
Other Info. Persuasion	0	0.00	0	0.00	0	0.00
Lyrical						
Song lyrics	457	1.86	70	0.28	527	2.15
Poem	31	0.13	23	0.09	54	0.22
Other Lyrical	0	0.00	2	0.01	2	0.01
Prayer	0	0.00	2	0.01	2	0.01
Spoken						
Interview	90	0.37	160	0.65	250	1.01
Transcript of video/audio	7	0.03	21	0.09	28	0.11
Formal speech	5	0.02	17	0.07	22	0.09
TV/movie script	0	0.00	12	0.05	12	0.05
Other Spoken	0	0.00	5	0.02	5	0.02
TOTAL	11,345	46.18	13,220	53.82	24,565	100.00

TABLE 10. Two-way hybrid categories that occurred more than 100 times, with frequency and percent information.

2-way hybrid	Frequency	% of 2-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
All other possible 2-2 coding splits (21)	470	8.3
TOTAL	5,682	100.0

TABLE 11. Three-way hybrid categories that occurred more than 100 times, with frequency and percent information.

3-way hybrid	Frequency	% of 3-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 possible 2-1-1 coding splits (44)	876	10.3
TOTAL	8,515	100.0

categories, we find that 72.5% of all hybrid documents were labeled as Informational Description/Explanation by at least one of the raters (i.e., 9,554/12,909). The register label of Opinion also appears in 12 of the hybrid categories, comprising ~70% of the hybrid documents. Finally, although the Narrative label only appeared in 10 of the 19 hybrid categories, it was selected by at least one of the raters for 71% of the hybrid documents. On the one hand, these frequency findings are unsurprising when we consider that Narrative, Informational Description/Explanation, and Opinion were the most frequent register categories used in the single register data reported above. However, it is interesting to note

that the characteristics of these general registers emerge in most hybrid texts as well as in texts that achieve majority agreement.

By considering these hybrid documents together with the documents that raters agreed on, we are able to provide an overall evaluation of our approach. Table 12 shows the percent of all web documents in our corpus that could be classified using the categories established above: 4-rater agreement, 3-rater agreement, 2-way hybrid, and 3-way hybrid. Thus, overall, when the hybrid texts are combined with the single category texts, we find that our methodological approach allowed us to classify over 95% of the corpus into register categories and more than 61% of the corpus into subregisters. However, it should again be emphasized that these hybrid registers and subregisters are exploratory at this stage, and additional evidence would be needed to determine whether they are valid text categories.

Up to this point in the paper we have focused on measuring the number of web documents that nonexpert users were able to classify into register, subregister, and hybrid categories. Another approach to assessing our data is to quantify user perceptions of the register categories in our framework. Essentially, this approach measures the likelihood that users will agree on a particular register category for a given text. Fleiss' kappa was calculated for each of the eight register categories to quantify user perceptions of how well defined a given register is (see Table 13). These results show a wide range of variation across register categories in terms of how well defined they are for nonexpert users. On the one hand, the Interactive Discussion and Lyrical register categories were very well defined for the users. This shows that (a) these text categories are clearly defined for most nonexpert users, and (b) texts in these categories have characteristics that make them distinguishable to most users.

On the other hand, the register of Information Persuasion was not well defined for the users. As we mention above, this category was developed during this study in order to make a distinction between texts that are primarily informational, but also have the intent to persuade, from texts that are primarily opinionated (see Table 3, step 9). However, this distinction does not seem to be clear in the minds of users. As a result, it was not used frequently and was seldom agreed upon by users.

The register categories with the second and third lowest Fleiss' kappa were Informational Description/Explanation and Opinion. Unlike the Informational Persuasion category, these two registers were among the most frequently chosen and agreed upon categories in our framework (see Table 8). While these results seem to contradict each other, they can be explained, at least in part, by the frequent occurrence of these two registers in the hybrid categories discussed above (see Tables 11 and 12).

## Discussion and Conclusions

The overarching goal of this study was to develop and assess a new user-based methodological approach for web

TABLE 12. Frequency, percent, and cumulative percent results for total number of texts classified.

	Register level			Subregister level		
	# of texts	% of total URLs	Cumulative %	# of texts	% of total URLs	Cumulative %
4-rater agreement	17,511	36.37	36.37	11,345	23.56	23.56
3-rater agreement	15,684	32.57	68.94	13,220	27.46	51.02
2-way hybrid (100+)	5,212	10.83	79.88	1,620	3.36	54.38
3-way hybrid (100+)	7,639	15.87	95.75	3,377	7.01	61.39
No agreement	2,101	4.23	100.00	18,585	38.61	100.00
Total	48,147	100.00	—	48,147	100.00	—

TABLE 13. Fleiss' Kappa coefficients indicating the extent of agreement between raters' perceptions of each register category.

Register category	Fleiss' kappa
Narrative	.51
Informational Description/Explanation	.37
Opinion	.36
Interactive Discussion	.86
How-to/Instructional	.47
Informational Persuasion	.26
Lyrical	.82
Spoken	.46

register classification. We first created and piloted a hierarchical register framework based on the situational characteristics of internet texts, which addresses many of the challenges of web register classification that have been identified by previous researchers. We implemented this framework in a user-based web register classification instrument and tested it through several rounds of piloting. After achieving high rates of reliability, we then recruited coders through MTurk to apply the instrument, in order to code a large corpus of ~50,000 web documents. Finally, we assessed the effectiveness of our methods according to several measures, providing strong support for the usefulness of our register framework as a tool for classifying internet texts into register, subregister, and hybrid categories.

One issue discussed in the previous literature is that not all register categories are equally meaningful to end users, and we encountered similar problems in our early rounds of pilot testing. To address this issue, we adopted a hierarchical approach, asking users to identify key situational characteristics of documents, rather than directly identifying specific registers and subregisters. This information was incorporated into our decision-tree survey, which guided raters to a final page with a short list of subregister options. The results show that users are able to reliably agree on register and subregister categories for the majority of documents. However, not all of these categories are equally well defined and, as a result, there was much less agreement regarding some register categories.

This method also made it possible to address the multifaceted nature of web registers, another major issue

identified by previous researchers. We identified several distinct situational parameters (or facets) that combine to capture the definition of a register category. Rather than asking users to assign a register label to a text by simultaneously accounting for the multiple parameters we had identified, we requested this information one parameter at a time, in a hierarchical fashion. This ultimately made it possible for the users to classify texts into complex, multifaceted categories by making a series of relatively simple decisions about the characteristics of the texts.

Previous researchers have also struggled to achieve the most appropriate level of register abstraction or specificity. The hierarchical nature of our register framework addresses this challenge by incorporating more than just one level of abstraction, in a hierarchical fashion. Additionally, by incorporating this same hierarchical structure into our decision-tree survey, we gain information from users about the situational characteristics of texts that allows us to classify texts and measure classification reliability characteristics of texts on several different levels.

Finally, two related challenges—fuzziness in taxonomies and hybrid texts—have been identified in previous research as issues that contribute to low user agreement. While some previous research has collected classification results from multiple users per text, these studies have not capitalized on these data in order to identify hybrid categories. The large number of documents classified in this study, each coded by four different raters, made it possible for us to use a bottom-up approach to identify hybrid categories. The systematic nature of these hybrid categories clearly supports the existence of hybrid texts on the internet.

### Limitations

One limitation of our study that should be noted is the origin of the internet corpus. In our study we were interested in describing the searchable web, in contrast with the entire content of the internet. Accordingly, we relied on Google to create a corpus, based on lists of URLs. Although this method has been used extensively in previous research, it should be noted that classifying the entire universe of documents on the World Wide Web was beyond the scope of our study. Thus, we have investigated the registers found on the “searchable web,” rather than the entire web.

1 Another limitation of this study is the loss of webpages  
2 due to the amount of time between the initial collection of  
3 URLs and the coding of register characteristics. We antici-  
4 pated some attrition during this time, but we did not expect  
5 the loss of nearly 7% of our data. A cursory analysis of those  
6 URLs did not reveal noticeable patterns in the unavailable  
7 webpages. Rather, this pattern apparently reflects the  
8 dynamic nature of the web, which is in flux to an even  
9 greater extent than we anticipated.

10 A final limitation worth noting here relates to the nature  
11 of our participant sample and data collection procedures.  
12 Overall, we were pleased with our decision to use MTurk to  
13 recruit participants. However, the large number of partici-  
14 pants and the nature of online data collection necessarily  
15 limited our ability to analyze the demographic characteris-  
16 tics of specific coders, or to monitor the performance of  
17 participants. We performed extensive training and a number  
18 of data screening procedures designed to raters who were  
19 not correctly completing the task. However, it is likely that at  
20 least some variability was introduced into our data through  
21 misunderstanding or inattentiveness on the part of individual  
22 coders.

### 23 *Future Research*

24 Overall, the approach we developed and apply here  
25 shows great potential for the classification of internet docu-  
26 ments into register categories. However, there are aspects  
27 of the methodology that could be revised in order to  
28 improve the reliability and generalizability of the register  
29 framework and classification instrument. Applications of  
30 the method with different corpora and participant samples  
31 will be needed to determine how to proceed with such  
32 improvements.

33 Future research on hybrid (sub)register categories is also  
34 needed. The hybrid categories in this study accounted for a  
35 large number of the documents in our corpus. We are cur-  
36 rently conducting detailed discourse analyses of the docu-  
37 ments that were classified as hybrids, revealing that these  
38 are in fact hybrid texts that serve multiple communicative  
39 purposes.

40 Finally, we are currently carrying out research on the  
41 lexico-grammatical characteristics of these different regis-  
42 ter and subregister categories. The eventual goals of  
43 this research are to provide a comprehensive linguistic  
44 description of register variation on the web, which can in  
45 turn be used as the basis for more accurate and robust  
46 automatic identification of the register category of web  
47 documents.

### 48 **Acknowledgments**

49 This material is based on work supported by the National  
50 Science Foundation under Grant No. 1147581. We thank  
51 Anna Gates and Rahel Oppliger for help with web document  
52 classification.

### 53 **References**

- 54 Atkinson, D., & Biber, D. (1994). Register: A review of empirical research.  
55 In D. Biber & E. Finegan (Eds.), *Sociolinguistic perspectives on register*.  
56 Oxford, UK: Oxford University Press, 351–385.
- 57 Baroni, M., & Bernardini, S. (2004). BootCaT: Bootstrapping corpora and  
58 terms from the web. *Proceedings of LREC 2004*, Lisbon: ELDA. pp.  
59 1313–1316.
- 60 Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2009). The  
61 WaCky wide web: A collection of very large linguistically processed  
62 web-crawled corpora. *Language Resources and Evaluation*, 43(3), 209–  
63 226.
- 64 Biber, D. (1995). *Dimensions of register variation*. Cambridge, UK: Cam-  
65 bridge University Press.
- 66 Biber, D., & Conrad, S. (2009). *Register, genre, and style*. Cambridge, UK:  
67 Cambridge University Press.
- 68 ~~Biber, D., & Egbert, J. (in review). Towards a user-based taxonomy of web  
69 registers.~~
- 70 Crowston, S., & Kwasnik, B. (2004). A framework for creating a faceted  
71 classification for genres: Addressing issues of multidimensionality. In  
72 *Proceedings of the 37th Hawaii International Conference on System  
73 Sciences*.
- 74 Crowston, S., Kwasnik, B., & Rubleske, J. (2010). Problems in the use-  
75 centered development of a taxonomy of web genres. In A., Mehler, S.,  
76 Sharoff, & M., Santini (Eds.), *Genres on the Web: Computational models  
77 and empirical studies*. New York: Springer.
- 78 Crowston, S., & Williams, M. (2000). Reproduced and emergent genres of  
79 communication on the World Wide Web. *The Information Society*, 16(3),  
80 201–216.
- 81 Egbert, J., & Biber, D. (2013). Developing a user-based method of web  
82 register classification. In S. Evert, E. Stemle, & P. Rayson (Eds.), *Pro-  
83 ceedings of the 8th Web as Corpus Workshop (WAC-8) @Corpus Lin-  
84 guistics*. pp. 16–23.
- 85 Haas, S., & Grams, E. (1998). Page and link classifications: Connecting  
86 diverse resources. *Proceedings of Digital Libraries Third ACM Confer-  
87 ence on Digital Libraries*. pp. 99–107.
- 88 Kwasnik, B., Crowston, K., Chun, Y.-L., D'Ignazio, J., & Rubleske, J.  
89 (2006). Challenges in creating a taxonomy for genres of digital docu-  
90 ments. *Proceedings of the Ninth International ISKO Conference*, 4–7  
91 July 2006, Vienna.
- 92 Landis, J., & Koch, G. (1977). The measurement of observer agreement for  
93 categorical data. *Biometrics*, 33, 159–174.
- 94 Lee, D. (2001). Genres, registers, text types, domains, and styles: Clarifying  
95 the concepts and navigating a path through the BNC jungle. *Language  
96 Learning and Technology*, 5, 37–72.
- 97 Paolacci, G., Chandler, J., & Ipeirotis, P.G. (2010). Running experiments on  
98 Amazon Mechanical Turk. *Judgment and Decision Making*, 5, 411–419.
- 99 Rehm, G., Santini, M., Mehler, A., Braslavski, P., Gleim, R., Stubbe, A.,  
100 Symonenko, S., Tavosanis, M., & Vidulin, V. (2008). Towards a refer-  
101 ence corpus of web genres for the evaluation of genre identification  
102 systems. In *Proceedings of the 6th Language Resources and Evaluation  
103 Conference*, Marrakech, Morocco. pp. 351–358.
- 104 Rosso, M.A. (2008). User-based identification of web genres. *Journal of the  
105 American Society for Information Science and Technology*, 59(7), 1053–  
106 1072.
- 107 Rosso, M.A., & Haas, S.W. (2010). Identification of web genres by user  
108 warrant. In A., Mehler, S., Sharoff, & M., Santini (Eds.), *Genres on the  
109 Web: Computational models and empirical studies*. New York: Springer.
- 110 Roussinov, D., Crowston, K., Nilan, M., Kwasnik, B., Cai, J., & Liu, X.  
111 (2001). Genre based navigation on the Web. *Proceedings of the 34th  
112 Annual Hawaii International Conference on System Sciences*, Digital  
113 Documents Track, IEEE Computer Science Press.
- 114 Santini, M. (2007). Characterizing genres of web pages: Genre hybridism  
115 and individualization. In *Proceedings of the 40th Hawaii International  
116 Conference on System Sciences (HICSS-40)*.
- 117 Santini, M. (2008). Zero, single, or multi? Genre of web pages through the  
118 users' perspective. *Information Processing and Management*, 44, 702–  
119 737.

1	Santini, M., & Sharoff, S. (2009). Web genre benchmark under construction. <i>Journal for Language Technology and Computational Linguistics</i> , 25(1), 125–141.	
2		
3		
4	Sharoff, S. (2005). Creating general-purpose corpora using automated search engine queries. In M. Baroni & S. Bernardini (Eds.), <i>WaCky! Working papers on the web as corpus</i> . Bologna, Italy: Gedit.	
5		
6		
7	Sharoff, S. (2006). Open-source corpora: Using the net to fish for linguistic data. <i>International Journal of Corpus Linguistics</i> , 11(4), 435–462.	
8		
	Sharoff, S., Wu, Z.K., & Markert (2010). The Web library of Babel: evaluating genre collections. In <i>Proceedings of the Seventh Language Resources and Evaluation Conference, LREC 2010, Malta</i> .	9
		10
		11
	Suri, S., & Watts, D.J. (2011). Cooperation and contagion in Web-based, networked public goods experiments. <i>PLoS One</i> , 6(3), e16836.	12
		13
	Vidulin, V., Luštrek, M., & Gams, M. (2009). Multi-label approaches to web genre identification. <i>Journal for Language Technology and Computational Linguistics</i> , 24(1), 97–114.	14
		15
		16











UNCORRECTED PROOF

## AUTHOR QUERY FORM

Dear Author,

During the preparation of your manuscript for publication, the questions listed below have arisen. Please attend to these matters and return this form with your proof.

Many thanks for your assistance.

Query References	Query	Remarks
1	*AUTHOR: Biber, 1988 has not been included in the Reference List, please supply full publication details.	
2	*AUTHOR: Biber & Finegan, 1994 has not been included in the Reference List, please supply full publication details.	
3	*AUTHOR: Lindeman & Littig, 2010 has not been included in the Reference List, please supply full publication details.	
4	*AUTHOR: Santini, 2010 has not been included in the Reference List, please supply full publication details.	
5	AUTHOR: Not in references.	
6	*AUTHOR: Biber et al., 1999 has not been included in the Reference List, please supply full publication details.	
7	AUTHOR: Please check all links to ensure they are working.	
8	AUTHOR: Please provide year, journal, and if possible pages.	
9	*AUTHOR: The * has not been mentioned in the table. Please cite the * in the relevant place in the table.	
10	*AUTHOR: "Box 6032" has been set after "Northern Arizona University" and "4071 JFSB" has been set after "Brigham Young University" Please confirm that this is correct.	

Note: The query which is preceded by \* is added by Toppan Best-set.

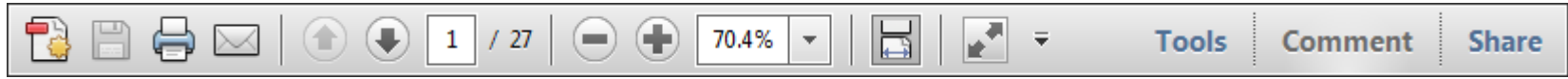


USING e-ANNOTATION TOOLS FOR ELECTRONIC PROOF CORRECTION

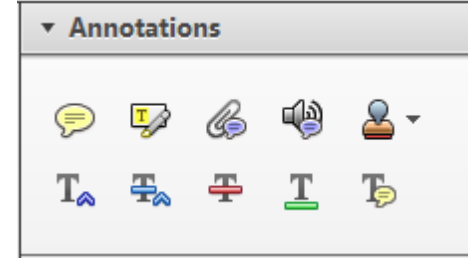
Required software to e-Annotate PDFs: Adobe Acrobat Professional or Adobe Reader (version 8.0 or above). (Note that this document uses screenshots from Adobe Reader X)

The latest version of Acrobat Reader can be downloaded for free at: <http://get.adobe.com/reader/>

Once you have Acrobat Reader open on your computer, click on the [Comment](#) tab at the right of the toolbar:



This will open up a panel down the right side of the document. The majority of tools you will use for annotating your proof will be in the [Annotations](#) section, pictured opposite. We've picked out some of these tools below:



**1. Replace (Ins) Tool – for replacing text.**



Strikes a line through text and opens up a text box where replacement text can be entered.

**How to use it**

- Highlight a word or sentence.
- Click on the [Replace \(Ins\)](#) icon in the Annotations section.
- Type the replacement text into the blue box that appears.

standard framework for the analysis of microeconomic activity. Nevertheless, it also led to the emergence of a number of strategic substitutes. The number of competitors in the industry is that the structure of the industry is a key determinant of the main components of the industry. At the industry level, are exogenous factors important? Works on entry by Shiraz (M henceforth) we open the 'black b



**2. Strikethrough (Del) Tool – for deleting text.**



Strikes a red line through text that is to be deleted.

**How to use it**

- Highlight a word or sentence.
- Click on the [Strikethrough \(Del\)](#) icon in the Annotations section.

there is no room for extra profits and the number of firms that can survive are zero and the number of firms (net) values are not determined by the number of firms. Blanchard and Kiyotaki (1987), perfect competition in general equilibrium. The effects of aggregate demand and supply in the classical framework assuming monopoly power. An exogenous number of firms

**3. Add note to text Tool – for highlighting a section to be changed to bold or italic.**



Highlights text in yellow and opens up a text box where comments can be entered.

**How to use it**

- Highlight the relevant section of text.
- Click on the [Add note to text](#) icon in the Annotations section.
- Type instruction on what should be changed regarding the text into the yellow box that appears.

dynamic responses of mark ups consistent with the VAR evidence

sation... y Ma... and... on n... to a... on... stent also with the demand-



**4. Add sticky note Tool – for making notes at specific points in the text.**



Marks a point in the proof where a comment needs to be highlighted.

**How to use it**

- Click on the [Add sticky note](#) icon in the Annotations section.
- Click at the point in the proof where the comment should be inserted.
- Type the comment into the yellow box that appears.

and supply shocks. Most of the... a... number... standard fr... cy. Nev... ole of st... ber of competitors and the imp... is that the structure of the secto



USING e-ANNOTATION TOOLS FOR ELECTRONIC PROOF CORRECTION

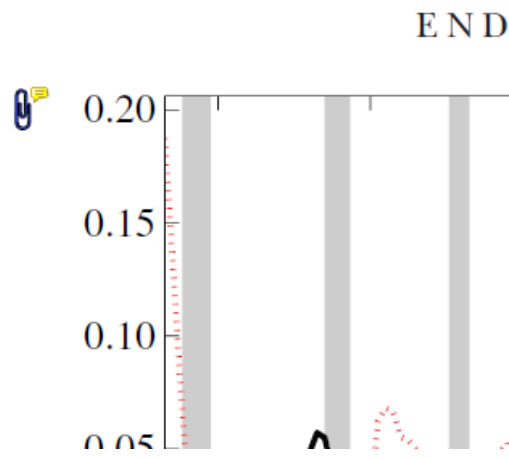
**5. Attach File Tool – for inserting large amounts of text or replacement figures.**



Inserts an icon linking to the attached file in the appropriate place in the text.

**How to use it**

- Click on the [Attach File](#) icon in the Annotations section.
- Click on the proof to where you'd like the attached file to be linked.
- Select the file to be attached from your computer or network.
- Select the colour and type of icon that will appear in the proof. Click OK.



**6. Add stamp Tool – for approving a proof if no corrections are required.**

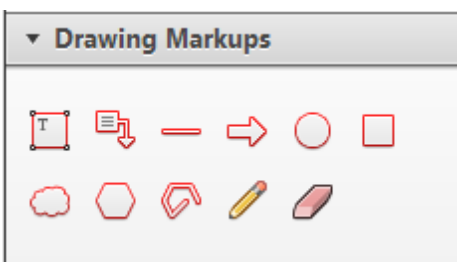


Inserts a selected stamp onto an appropriate place in the proof.

**How to use it**

- Click on the [Add stamp](#) icon in the Annotations section.
- Select the stamp you want to use. (The [Approved](#) stamp is usually available directly in the menu that appears).
- Click on the proof where you'd like the stamp to appear. (Where a proof is to be approved as it is, this would normally be on the first page).

of the business cycle, starting with the  
 on perfect competition, constant ret  
 production. In this environment goods  
 extra profits and the market for marke  
 he market for goods is determined by the model. The New-Key  
 otaki (1987), has introduced produc  
 general equilibrium models with nomin  
 and market-clearing. Most of this literat

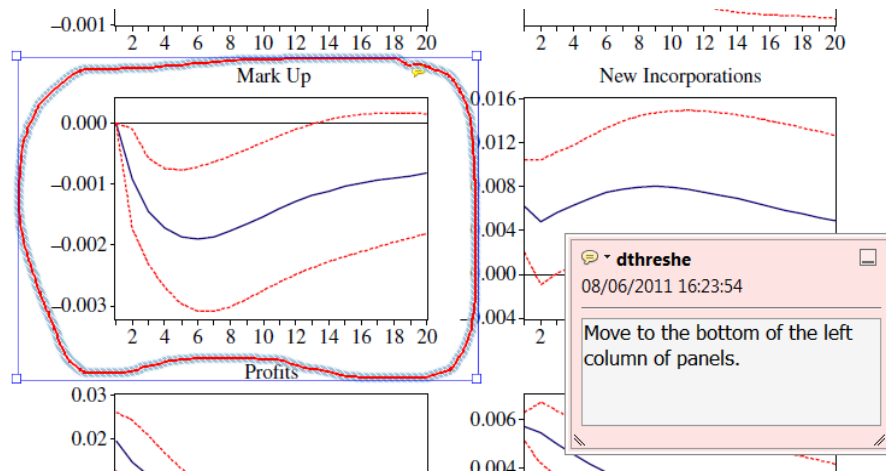


**7. Drawing Markups Tools – for drawing shapes, lines and freeform annotations on proofs and commenting on these marks.**

Allows shapes, lines and freeform annotations to be drawn on proofs and for comment to be made on these marks..

**How to use it**

- Click on one of the shapes in the [Drawing Markups](#) section.
- Click on the proof at the relevant point and draw the selected shape with the cursor.
- To add a comment to the drawn shape, move the cursor over the shape until an arrowhead appears.
- Double click on the shape and type any text in the red box that appears.



For further information on how to annotate proofs, click on the [Help](#) menu to reveal a list of further options:

