Developing a User-based Method of Web Register Classification

Jesse Egbert

Douglas Biber

Northern Arizona University

Introduction

- The internet has tremendous potential for linguistic research and NLP applications
- One central issue: Identification of web register/genre and analysis of register variation
- Previous research:
  - classification by experts and non-expert internet users
  - attempts to classify internet texts using Automatic Genre Identification (AGI)
- The present study: Extend previous research based on the perceptions of a large representative sample of internet users

Expert and user-based identification

- Classification of internet texts has been attempted by web genre experts and non-expert internet users
- However, inter-rater reliability among web genre experts tends to be quite low, especially for randomly extracted web texts (Sharoff et al., 2010)
- Non-expert users also vary in their understanding of genre labels (Crowston, Kwasnik, & Rubleske, 2010), and reliability among users is often unacceptably low (Rosso & Haas, 2010)

Automatic Genre Identification

- Some AGI approaches have achieved high accuracy rates (Sharoff, Wu, & Markert, 2010)
- However, past AGI research has some potential limitations:
  - we often don’t know whether our web corpora are representative (Santini & Sharoff, 2009)
  - **more importantly, we don’t know if the categories we are predicting are valid**
    - corpora are typically sub-divided into genre classes by only one person (Sharoff et al., 2010)

Research questions

- In order to address these research gaps, we set out to answer the following questions:

1. What are the web register distinctions recognized by non-expert internet users?

2. To what extent can non-expert raters reliably classify web texts into those register categories?

3. What is the distribution of English language registers on the web?

Corpus construction

- Mark Davies constructed a large corpus of internet language from 20 English-speaking countries (c. 1.9 billion words; 1.8 million web pages)*
  - URLs collected from results of Google searches of frequent English 3-grams (see Baroni & Bernardini, 2004; Baroni et al., 2009; Sharoff, 2005; 2006)
  - For this project we randomly extracted URLs from a subset of this corpus (US, UK, CA, AU, NZ)

*Davies' **Corpus of Global Web-based English (GloWbE)** is now freely available at http://corpus2.byu.edu/glowbe/

Development of the web register taxonomy

- Reviewed a large number of studies that proposed web register/genre palettes
- Began with 78 categories from the wiki-based collaboration on webgenrewiki.org
- Grouped these categories into 8 general registers (e.g., opinion, non-fiction narrative)
- Each general register contained several sub-register categories (opinion: opinion blogs, editorials, reviews, advice)

The final taxonomy

General Register	Sub-register examples
Narrative	News report/blog; personal/diary blog
Opinion	Opinion blog; review
Description	description of a person; research article
Discussion	Question/answer forum; other forum
Lyrical	Song lyrics; poem
How-to/Instructional	How-to; technical support
Informational Persuasion	Description with intention to sell; persuasive article/essay
Spoken	Interview; formal speech

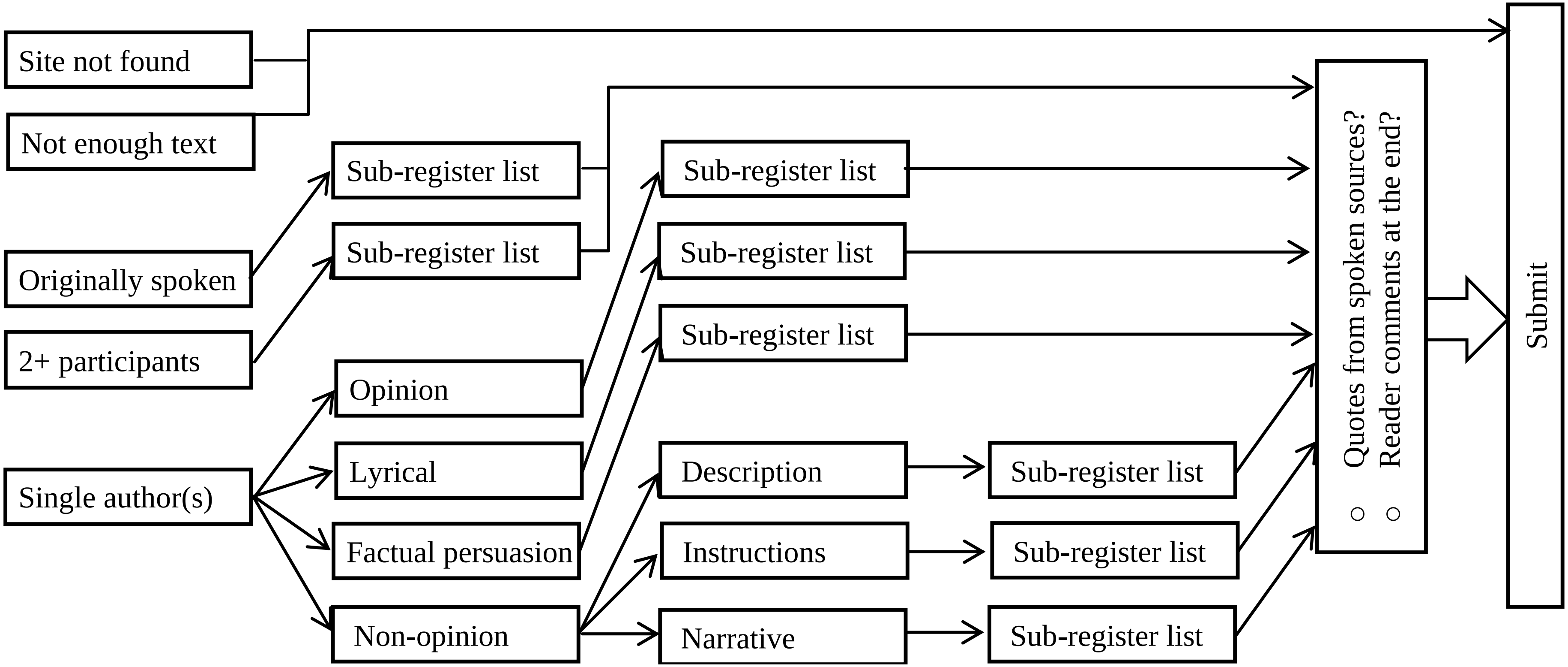
Instrument development

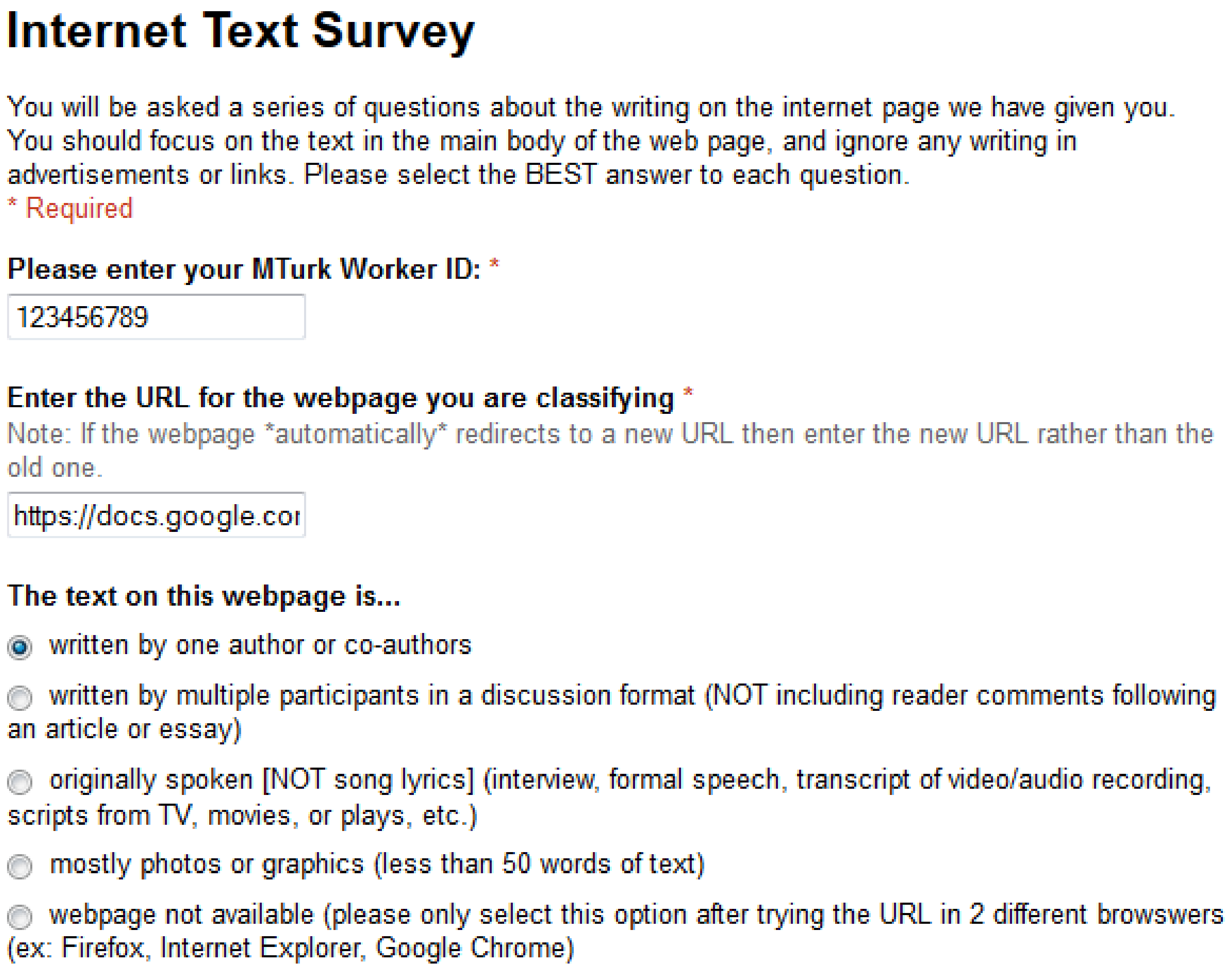
- Used a series of ten pilot studies to develop and refine an instrument capable of measuring register distinctions that:
  - are recognized by end-users
  - can be applied by end-users with high reliability
- Instrument developed in three stages: (1) rubric with descriptions; (2) flowchart with examples; (3) computer-adaptive online survey
- After each pilot study inter-rater agreement was measured and improvements were made

A new approach to user-based web register classification

- Computer-adaptive Google Form survey
- Raters guided through a series of 2-6 pages
- Each page contained a set of multiple choice options regarding the situational characteristics of a web text, such as:
  - The main purpose of this text is to…
    - narrate or report on PAST EVENTS
    - describe or explain INFORMATION
    - explain HOW-TO or INSTRUCTIONS
- Rater responses were used to classify texts into general register and sub-register categories
- On the final page raters were asked to check a box if the text contained reader comments

Schema for the final classification instrument





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: *

123456789

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.

https://docs.google.cor

The text on this webpage is...

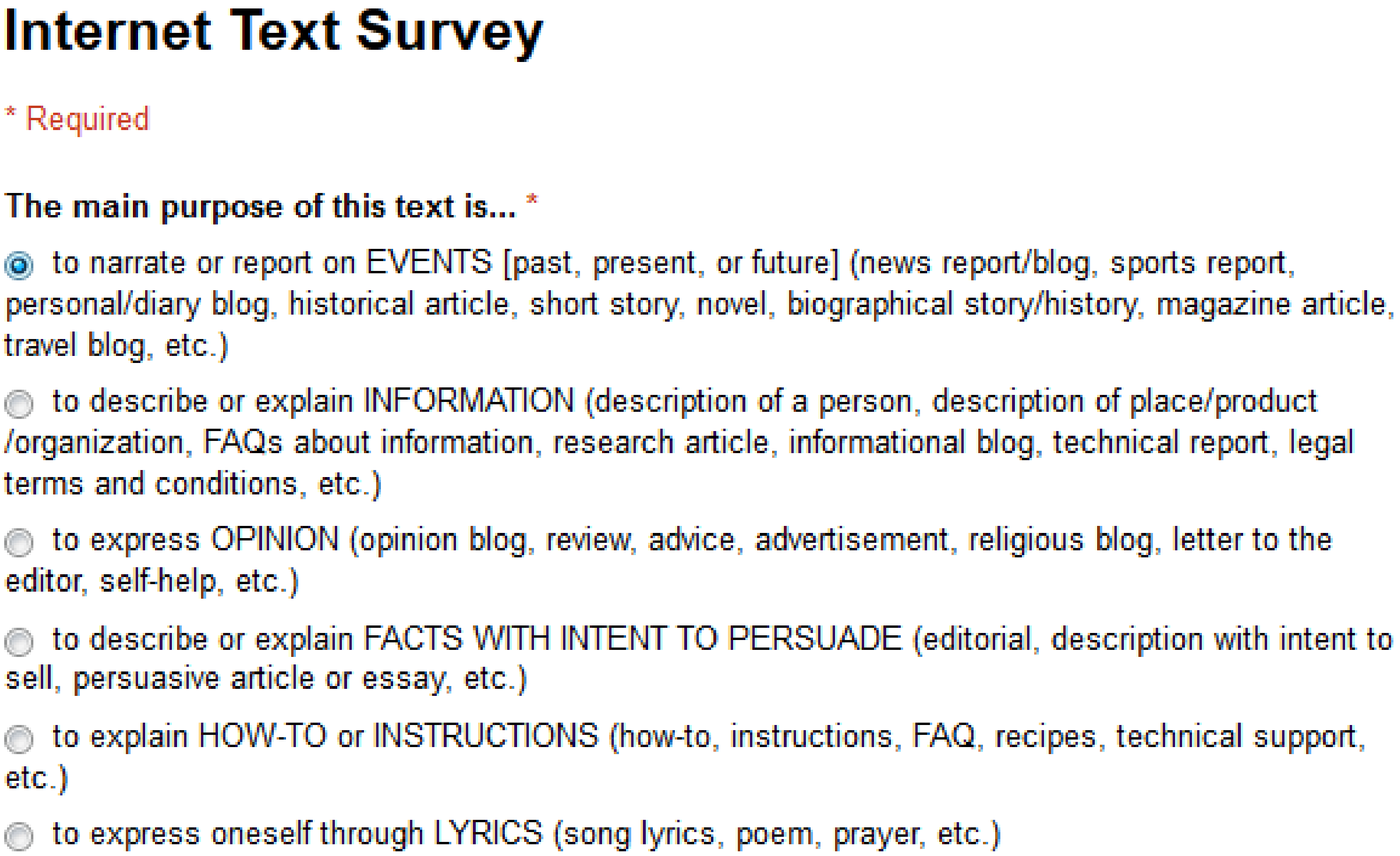
● written by one author or co-authors

○ 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 browswers (ex: Firefox, Internet Explorer, Google Chrome)



Internet Text Survey

* Required

The main purpose of this text is... *

● to narrate or report on EVENTS [past, present, or future] (news report/blog, sports report, personal/diary blog, historical article, short story, novel, biographical story/history, magazine article, travel blog, etc.)

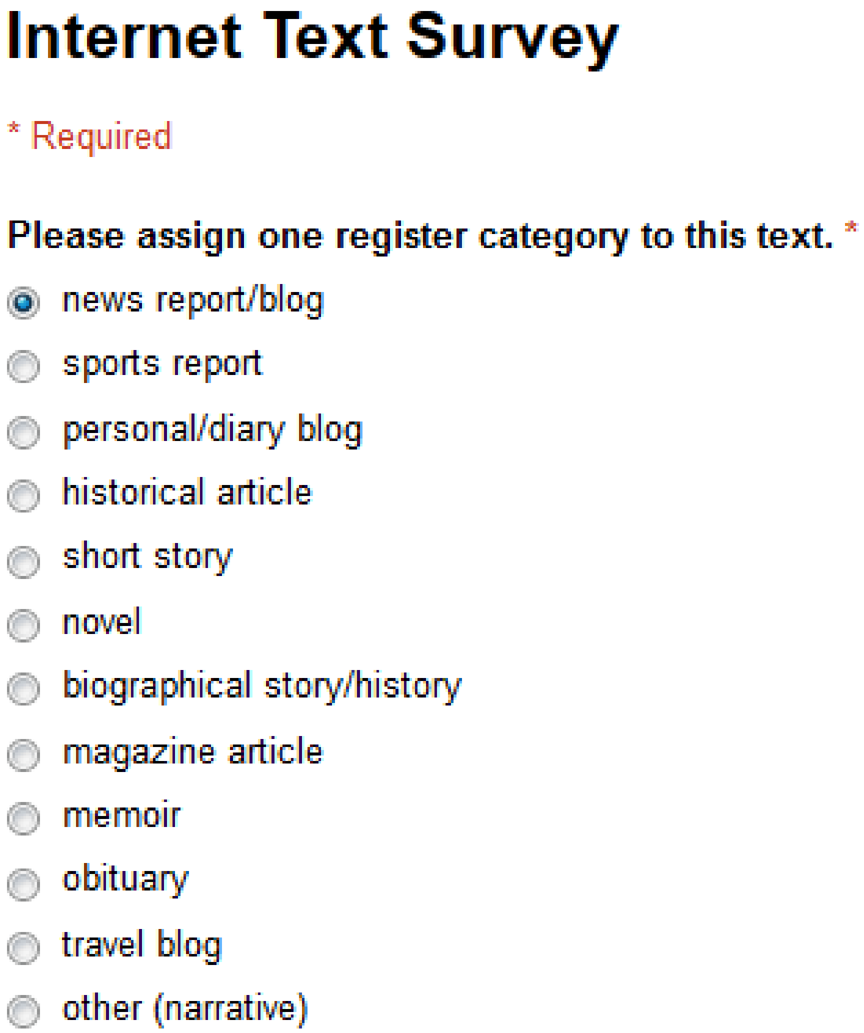
○ to describe or explain INFORMATION (description of a person, description of place/product /organization, FAQs about information, research article, informational blog, technical report, legal terms and conditions, etc.)

○ to express OPINION (opinion blog, review, advice, advertisement, religious blog, letter to the editor, self-help, etc.)

○ to describe or explain FACTS WITH INTENT TO PERSUADE (editorial, description with intent to sell, persuasive article or essay, etc.)

○ to explain HOW-TO or INSTRUCTIONS (how-to, instructions, FAQ, recipes, technical support, etc.)

○ to express oneself through LYRICS (song lyrics, poem, prayer, etc.)

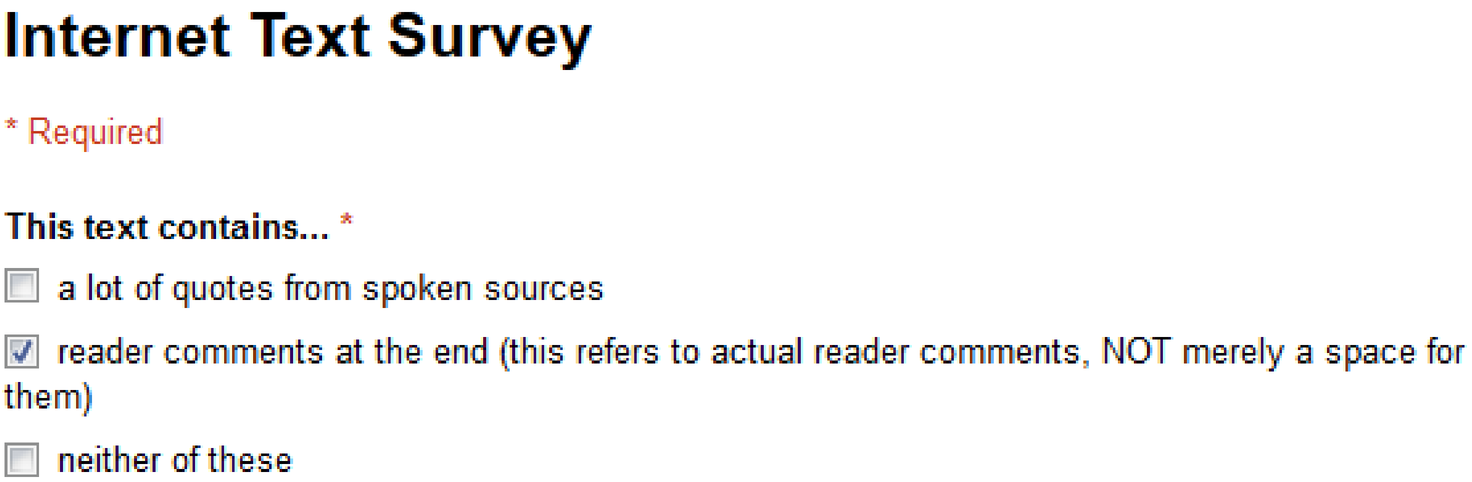


Internet Text Survey

* Required

Please assign one register category to this text. *

● news report/blog
○ sports report
○ personal/diary blog
○ historical article
○ short story
○ novel
○ biographical story/history
○ magazine article
○ memoir
○ obituary
○ travel blog
○ other (narrative)



Internet Text Survey

* Required

This text contains... *

□ a lot of quotes from spoken sources

☑ reader comments at the end (this refers to actual reader comments, NOT merely a space for them)

□ neither of these

Final pilot

- 1,000 URLs were randomly selected from the corpus database
- Raters were recruited through Mechanical Turk, an Amazon crowdsourcing company
- 4 different people rated each of the web texts
- A single register category was assigned if at least 3 of the 4 raters agreed
- Additionally, we allowed for ‘hybrid registers’
  - Operational definition: 2-2 and 2-1-1 ties

Final pilot II

- 3.6% of the URLs were not found (site down, page removed, broken link, etc.)
- 3.3% of the texts were labeled as not having enough text to rate (less than 50 words of running text)
- Final dataset contains 931 webpages

Agreement results

General Registers				
4 agree	3 agree	2-2 hybrid	2-1-1 hybrid	No agreement
315	269	104	173	70
33.8%	28.9%	11.1%	18.6%	7.6%
Sub-registers				
4 agree	3 agree	2-2 hybrid	2-1-1 hybrid	No agreement
171	231	73	90	366
18.3%	24.8%	7.8%	9.8%	39.3%

General register distribution

General Register	#	%
Narrative	135	33.6
Opinion	95	23.6
Description	67	16.7
Discussion	54	13.4
Lyrical	18	4.5
How-to/Instructional	16	4.0
Informational Persuasion	10	2.5
Spoken	7	1.7

Sub-register distribution

Register	#	%
Narrative	135	
News report/blog	99	73.3
Sports report	19	14.1
Personal/diary blog	7	5.2
Historical article	4	3.0
Short story	3	2.2
Novel	2	1.5
Biographical story/history	1	0.07
Joke	0	0
Magazine article	0	0
Memoir	0	0
Obituary	0	0
Other factual narrative	0	0
Other fictional narrative	0	0
Other personal narrative	0	0
Travel blog	0	0
Opinion	95	
Opinion blog	57	60.0
Review	23	24.2
Advice	9	9.5
Religious blog/sermon	5	5.3
Self-help	1	1.1
Advertisement	0	0
Letter to the editor	0	0

Sub-register distribution

Register	#	%
Description	67	
Description of a thing	34	50.7
Description of a person	9	13.4
Research article	7	10.4
Abstract	5	7.5
Legal terms and conditions	4	6.0
FAQ about information	2	3.0
Encyclopedia article	2	3.0
Informational blog	2	3.0
Course materials	1	1.5
Technical report	1	1.5
Other	0	0
Discussion	54	
Question/answer forum	46	85.2
Other forum	7	13.0
Other discussion	1	1.8
Reader/viewer responses	0	0
Lyrical	18	
Song lyrics	17	94.4
Other	1	5.6
Poem	0	0
Prayer	0	0

Sub-register distribution

Register	#	%
How-to/Instructional	16	
How-to	13	81.3
Technical support	2	12.5
Recipe	1	6.2
Instructions	0	0
FAQ about how to do something	0	0
Other	0	0
Informational Persuasion	10	
Description with intent to sell	8	80.0
Persuasive article or essay	2	20.0
Editorial	0	0
Other	0	0
Spoken	7	
Interview	5	71.4
Formal speech	1	14.3
Transcript of video/audio	1	14.3
Other	0	0
TV/movie script	0	0

Frequent hybrid combinations (2+2)

Hybrid Combination (2+2)	Count
Description + Narrative	43
Narrative + Opinion	27
Description + Opinion	17
Informational Persuasion + Opinion	11
Description + Informational Persuasion	6

Frequent hybrid combinations (2+1+1)

Hybrid Combination (2+1+1)	Count
Narrative + Opinion + Description	56
Description + Informational Persuasion + Opinion	40
Description + Informational Persuasion + Narrative	28
Informational Persuasion + Narrative + Opinion	24
Description + How-to/Instructional + Opinion	15

Distribution of web pages with reader comments

Register	Count	Percent
Narrative	87	37.2
Opinion	86	36.8
Description	37	15.8
Informational Persuasion	12	5.1
How-to/Instructional	8	3.4
Lyrical	4	1.7
Spoken	0	0
Discussion	0	0
Total	234	100

Discussion

- The majority of internet texts can be reliably classified into web registers by non-expert internet users
- General register categories:
  - Majority agreed for 62.7% of texts
  - 30% of webpages were classified as hybrids
  - Over 92% of webpages were classified into meaningful categories
- Sub-register categories:
  - Majority agreed for 43% of texts
  - 17.5% of webpages were classified as hybrids
  - 61% of webpages were classified into meaningful categories

Discussion

- There is a great deal of register variation on the internet
  - 35/56 sub-register categories were agreed on for at least one text
- However, a relatively small number of general registers and sub-registers account for a large proportion of internet texts
  - 87% of all webpages were classified into one of the four most frequent general register categories (Narrative, Opinion, Description, Discussion)
  - More than half of all texts were classified into one of the three most frequent sub-registers (News report/blog, Opinion blog, Question/answer forum)

Discussion

- Many internet texts have the characteristics of more than one register category
- Our study was the first to use a bottom-up approach to empirically identify web register hybrids in a large-scale study
- Future research will be needed to fully understand web register hybrids, but the approach used in this study seems to be a viable approach

Discussion

- One of the most important attributes of language on the internet is the potential for interactivity among multiple participants
  - More than a quarter of all webpages in our final pilot analysis contained reader comments

Future steps

- Use the web register classification survey developed here to classify 50,000 random webpages
- Complete comprehensive linguistic descriptions of all 50,000 documents
- Determine whether the results of the linguistic analysis can be used to accurately predict the register of a web text
- Automatically apply the register framework to a 100 million word web corpus
- Make the corpus available in tagged and register-annotated form through Mark Davies' web-based corpus interface

Thank you

Questions?

Jesse.Egbert@nau.edu
Douglas.Biber@nau.edu
Mark_Davies@byu.edu