Towards a reference corpus of web genres

Marina Santini[1] and Serge Sharoff[2]

[1]University of Brighton
[2]University of Leeds

Rationale for the Colloquium

Genres of spoken and written texts are being intensively studied from various angles, e.g., communication studies, discourse analysis, computational linguistics, without arriving at a generally accepted definition. Many corpora have been built to represent the language, but very few large corpora indicate genres, and when they do the typology of genres varies widely. For instance, the Brown corpus famously uses 15 textual categories, from press reportage (a text genre) to religion or skills and hobbies (domains), while the British National Corpus (BNC) uses 70 classes, such as academic or non-academic scientific texts or biography. Interestingly, genre classes in the BNC are an add-on proposed by David Lee (2001)[1] after the corpus construction, rather than a basic criterion of the corpus creation. The genre attribute was included in a few collections used in Information Retrieval (TREC HARD 2003 and TREC HARD 2004). More precisely, in TREC HARD 2003, genre was included among the metadata in the following form:

item=GENRE represents the type of material the searcher is interested in.

- value=OVERVIEW means the searcher is interested in general news related to the topic.
- value=REACTION indicates the searcher is looking for news commentary on the topic.
- value=I-REACTION is like REACTION but is specifically about non-U.S. news commentary.
- value=ADMINISTRATIVE means the search is interested in official US government documents.
- value=ANY indicates that any genre is acceptable or none was indicated.

Instead, in TREC HARD 2004 genre had values of NEWS-REPORT, OPINION-EDITORIAL, *other*, or *any*. In TREC HARD 2005 the genre attribute was not included.

The web is new, so it is even less not clear how to apply traditional notions of genre to web pages.

In corpus-based genre studies, the main tendency has been to build one's own genre collection according to subjective criteria for corpus composition, genre annotation, and genre granularity. This is especially true for genre studies based on collections of web pages. What is more, no genre annotation criteria have been explicitly agreed upon. Genre annotation has been based either on the common sense of a single rater, or on the

[1]Lee D. (2001). "Genres, Registers, Text types, Domains, and Styles: Clarifying the concepts and navigating a path through the BNC Jungle". *Language Learning & Technology*. Vol. 5, No. 3, pp. 37-72.

agreement of few annotators. In brief, as it is now, web genre analyses remain self-contained and corpus-dependent.

Building a reference corpus of web genres is certainly difficult because web documents are often characterised by a high level of genre hybridism, by a fragmentation of textuality across several documents, by the impact of technical features such as hyperlinking, posting facilities and multi-authoring. Since the web is a huge reservoir of documents that can be easily mined for building all sorts of corpora, it is important to overcome the subjectivity that characterizes genre-related issues, in order to create sharable resources. What should we consider when designing a reference corpus of web genres? Genres of web pages show some traits that are not accounted for in TREC collections or in the BNC and that are, instead, important on the web. For example:

- *Genre Hybridism and Individualization*
  The fluidity and fast-paced dynamism of the web together with the complexity of web pages cause unclear genre conventions, and favour genre mixture and authorial creativity. These two phenomena appear to be very common on the web.

- *Granularity of the Unit of Analysis*
  How many granularities of the unit of analysis should be included? Only genres representing web sites? Only genre representing web pages? Both?

- *Format of Web Documents*
  An issue related to the previous one is represented by the 'format' that should be used to store the 'units of analysis' in a collection. In what form can a web page or a website be included in a corpus? In HTML format or in a text-only version? Including images or leaving them out? Removing boilerplates or keeping them? In, a database-like form, as DOM trees, as a net of graphs, in HTML format, or simply in a text-only version?

- *Genre Granularity and Similarity*
  Genres can be accounted for at subgenre, genre and super-genre level: what level of genre granularity and similarity should be applied in the reference corpus? Furthermore, should similar genres, such as TUTORIAL and HOW-TO, be accounted for separately?

- *How to build a Genre Palette*
  How many and which genres should be included in a genre reference corpus?

- *Validation and Evaluation of a Reference Corpus of Web Genres*
  How can we validate and evaluate the quality of a genre corpus?

The *rationale* for this colloquium is to draw up an initial list of characteristics and requirements for building, annotating and evaluating reference corpora of web genres.

Four longer presentations prepared for the colloquium report empirical results and offer hands-on answers to some of these questions. More precisely, Alexander Mehler analyses web genres at website level and suggests a database-like form of storage. He offers an interesting angle on the notion of web genres using structural and linking information. Barbara H. Kwaśnik, Kevin Crowston, Joseph Rubleske, You-Lee Chun tell us how they built a corpus of genre-tagged web pages to populate their genre collection. Serge Sharoff focuses on the similarities between web-derived corpora and classical corpora constructed from print media. Finally, Mark Rosso describes his

experience in assembling a genre palette that could be useful for building a genre reference corpus to help web searches.

Shorter presentations describe settings of ongoing or future research, and provide preliminary answers to some of the problems listed above. More precisely, Andrea Stubbe and Christoph Ringlstetter discuss two important aspects in web genre research: granularity of genre hierarchies and multi-genre classification. Rosario Caballero and Noelia Ruiz-Madrid focuses on HOW-TO TEXTS and address the issue of similar genres. Andrea Stubbe, Christoph Ringlstetter, Tong Zheng, and Randy Goebe present an intriguing idea: a genre classifier that adapts to the information need of a specific user on the basis of user events. They report on how to assemble a genre-annotated corpus. Julia Almeida points out the importance of the pictorial information, which currently plays a miner role in genre analysis and corpus building and which might deserve more attention when dealing with web documents. Finally, Cornelius Puschmann proposes an XML-based storage schema for the compilation of computer-mediated discourse (CMD) corpora from mixed sources.

In conclusion, building a genre-annotated reference corpus of web pages is arduous for a number of reasons, and several solutions appear to be viable. In this colloquium, we would like to make a first attempt to apply the concept of genre to the development of sharable criteria for building genre corpora. The ambition of this colloquium, the first ever organized on this topic, is to bring together researchers from different communities such as corpus linguistics, genre analysis, digital genre community, computational linguistics, and information retrieval in order to promote the discussion and development of new ideas and methods to create new corpora for language studies and as evaluation resources.

Authors, Titles and Short Abstracts of the Presentations

Longer Presentations

Alexander Mehler: *A Corpus Model of Structure Formation in Hypertext Types*

> This paper describes a web genre corpus model. Its starting point is a graph model of the logical document structure of hypertext types and of the linkage of their constituents. We describe an XML-based serialization of this model and provide a database mapping which retains a wide range of web genre data. This will be exemplified by three web genres.

Barbara H. Kwaśnik, Kevin Crowston, Joseph Rubleske and You-Lee Chun: *Building a Corpus of Genre-Tagged Webpages for an Information-Access Experiment*

> This presentation reports on one phase of a larger study whose overarching aim is to determine how providing genre metadata can help in access to sources of information in a digital environment. We have built a corpus of genre-tagged web pages and structured this particular experimental corpus in such a way as to provide the maximum control for our experiments. We recognize, however, that much rich genre information was either too difficult to represent or had to be pared away.

Serge Sharoff: *In the garden and in the jungle: comparing genres in the BNC and Internet*

> According to Adam Kilgarriff the BNC is a jungle when compared to smaller Brown-type corpora, but it looks more like an English garden when compared to the Internet. In this presentation I will compare English and Russian Internet corpora against their human-

collected counterparts (BNC and RNC) using two methods: the first involves manual annotation of a subset of Internet corpora, the second one uses probabilistic classifiers. The study shows that the Internet is not radically different from the BNC: Internet corpora do contain a wide range of genres and approximate many genres that exist in their printed form, the same is true for the audience level (texts for professional or layman texts).

Mark Rosso: *Development of a Genre Palette*

This presentation details the development of a genre palette used in the study of the effects of genre-annotated search results on the relevance judgement process in a web search environment. This palette development was conducted in several phases: (i) a survey of user terminology; (ii) user-based refinement of terminology into a tentative genre palette, and (iii) user validation of the genre palette.

Shorter Presentations

Andrea Stubbe and Christoph Ringlstetter: *Recognizing Genres*

We introduce a two-level hierarchy of genres based on the definition of genre in terms of form and function (or purpose). Thereby we provide sufficient granularity with the possibility to return to a coarser scheme when preferable. As some texts may naturally fall into more than one genre, an assignment to multiple classes is possible. For those applications where a unique class is required, several techniques for the combination of classifiers were evaluated.

Rosario Caballero and Noelia Ruiz-Madrid: *The impact of technology on how-to texts: Issues and prospects*

This paper explores online how-to texts produced by private and public entities. Together with analysing the link system of the texts (using C-map) we discuss (a) whether authorial differences have an impact on the texts' architecture, (b) the way(s) users search for the texts on the web and their representation of the genre, and (c) the heading used to store such a corpus.

Andrea Stubbe, Christoph Ringlstetter, Tong Zheng, and Randy Goebe: *Incremental genre classification*

In this presentation we will describe attempts to acquire data. These attempts have to consider the users explicitly and cooperatively. The user behaviour will be simulated using annotated corpus data. We will also formulate different scenarios for information gain representing different levels of uncertainty. Our goal is to integrate existing material of different sources into a realistic application.

Julia Almeida: *Text and image in web context*

We will propose new connections between linguistics and semiotics in order to redefine the relations between image and text. We intend to construct an approach to elucidate peculiarities of a texts corpus from web (including several genres) and explore the notion of textuality in web context.

Cornelius Puschmann: *SchemaCMD: An XML-based storage schema for the compilation of mixed-source CMD corpora*

This presentation will outline an XML schema for the segmentation and storage of data from Internet sources, specifically those which utilize so-called web feeds (often associated with the RSS protocol). It is based on the faceted classification scheme recently proposed by Susan Herring and aims to make data from diverse sources accessible and comparable in a single format.

Organization of the Colloquium

The colloquium lasts approximately 4.5 hours and is organised as follows:

- Opening (10 minutes)
- 4 longer presentations, i.e. 30-min presentations: including 20-min talk + 10 min for discussion (30 min * 4 = 120 min);
- Break (20 min)
- 5 shorter presentation, i.e 20-min presentations: including 15-min talk + 5 min for discussion (20 min * 5 = 100 min);
- Final discussion and winding up (30 minutes)

Grand total: approximately 4.5 hours (280 min)

Organizers:

Marina Santini, University of Brighton (Marina.Santini@itri.brighton.ac.uk)

Serge Sharoff, University of Leeds (s.sharoff@leeds.ac.uk)

~~*~*