Building Reference Corpora of Web Genres: An Initial List of Specifications

The rationale of this Colloquium is to draw up an initial list of specifications for building, annotating and evaluating reference corpora of web genres.

The main aim is to create sharable and reliable resources, and encourage their re-use and enhancement, without starting from scratch every time a genre collection of web documents is needed. The ultimate goal is to create a repository where all these resources are documented and can be freely accessed, used, improved and augmented.

What do we wish to know when using or creating a web corpus annotated by genre? Corpus specifications account for the level of generality we can attribute to our findings. Therefore, it is important to identify what is worth knowing about a corpus of genres downloaded from the web. The web is an important source for the creation of corpora, and genre is an important dimension that provides insights into language and texts. What should we know when using or creating a web corpus annotated by genre? How can we can make our findings comparable and how can we create sharable resources? In brief, this Colloquium is an attempt to establish some kind of commonality, i.e. a common ground for discussion about web-derived genre-annotated resources.

The following list is just an initial breakdown of the type of details we would like to know about a corpus of web genres.

The specifications listed below are based on **external criteria**. This means they are descriptors, i.e. they describe the decisions made by the creator(s) of a web genre corpus. The Colloquium presenters kindly helped us fill up these specifications for demonstrative purposes.

What about **internal criteria**, i.e. criteria that can describe web genres drawing from the documents themselves? For example, could it be that internal criteria are more suitable to measure corpus representativeness or intra-genre variation than external criteria? We do not know yet.

Ideally, external and internal criteria should complement each other in order to give a more comprehensive description of a corpus of web genres.

Comments, suggestions, proposals, projects, plans for future directions, positive and negative criticism are highly appreciated. Please send them to Marina Santini (***MarinaSantini.MS@gmail.com***) and Serge Sharoff (***s.sharoff@leeds.ac.uk***).

THE INITIAL LIST OF SPECIFICATIONS INCLUDES:

1) Purpose of genre annotation (e.g. to study language in context, for lexicographic purposes, for automatic genre classification, for information extraction)
2) Motivation and construction of the genre palette (e.g. why are the genres included in the palette interesting?)
3) Granularity of web documents (e.g. web pages, web sites, web page body, paragraphs)
4) Granularity and number of genre classes (e.g. super-genres, genres, subgenres, neighbouring categories)
5) Similarity of genre classes or fuzzy genre labels (e.g. online tutorials vs. users' manuals (similar), or online tutorials vs. blogs (dissimilar)

6) Genre classification scheme (e.g. one genre per web page, multiple genres per web page; multiple genre per website but a single genre per web page);

7) Genre features (e.g. links, POSs, content words, HTML tags)

8) Corpus format (e.g. XML (trees), HTML files (only tags), text only files, HTML including pictures)

9) Corpus storage (database, XML, text only, snippets)

10) Corpus construction and annotation (e.g. how were web documents selected, how many annotators?)

11) Size of the genre-annotated corpus and genre distribution (how big is the corpus annotated by genre, how many documents per genres)

12) Corpus evaluation and corpus representativeness: how do we assess the reliability of a genre corpus? How do we know that the corpus does not overfit the purpose? (e.g. with multiple raters (how many? 3, 5, 20, 50, 100?; with genre analysts' annotation? annotation by the web document creators?; with statistical measures?)

FILLING OUT THE SPECIFICATIONS

Our speakers have helped us show how these specifications can be filled up by presenting their work in this Colloquium. Therefore now we know the details about existing genre collections, their motivation, their scope.

1) Purpose of genre annotation

- **Puschmann: to carry out a contrastive analysis of individual language production and to correlate the results with the degree of stylistic intra-genre variation. One central question is: how consistent is a presupposed genre in terms of style?**
- **Ringlstetter (Incremental genre classification): to study the search interface and user behaviour. To improve search.**
- **Rosso: to build a genre palette for improving the effectiveness of web searching.**
- **Rubleske:** to provide genre metadata to help access information in a digital environment. Corpus built for a controlled experiment.
- **Sharoff: to annotate large diverse corpora**
- **Stubbe (Recognizing genres): automatic genre classification (supervised machine learning).**

2) Motivation and construction of the genre palette

- **Puschmann: Motivation. The study of an emerging genre: the *corporate blog*. Construction. Subjective selection of the author[?].The selection is not subjective in the sense that the author includes all blogs available that meet his central definition, i.e. a corporate blog is any blog that is maintained by the employee of a corporation that is used to facilitate organizational goals. Other criteria are language (the blog must be in English) and size of the organization (the vast majority of companies that he has included are publicly traded and have more than 10 employees). The idea of specifically looking at corporate blogs (and not other types of writing) is original, and is put forward by the author for the first time. Note that the corporate blog is**

a very young genre that is still emerging, i.e. the number of blogs of this type is quite limited.

- **Ringlstetter: Motivation: Study the impact of user feedback. Construction, Re-use of an existing web genre collection.**
- **Rosso: Motivation. To find genre categories that enjoy widespread recognised by their intended users groups. Construction. 3 user studies. The first study was a survey of user terminology for web pages. The second study aimed to refine the resulting set of forty-eight (often conceptually and lexically similar) genre names and definitions into a smaller palette of user-preferred terminology. The third study aimed to show that users would agree on the genres of web pages, when choosing from the genre palette.**
- **Rubleske:** Motivation. To develop a 'palette' that is useful to people performing certain types of web search tasks. Construction. While performing actual web searches, study participants from three knowledge domains (primary education, print journalism, aerospace engineering) offered verbal descriptions of the web pages they viewed (they were asked to state the page's 'type'). The palette constructed for the follow-up study is based on the taxonomy generated from this data and augmented by an analyst on the research team. To meet the needs of the experiment, terms from the palette were organized into a shallow hierarchy.
- **Sharoff: Motivation. As the use of a list of genre labels is problematic (five good reasons), he suggests the use of 'communicative intentions' underlying the creation of texts in respective genres, complemented with other classification dimensions, such as mode or audience. Construction. Initial ideas of John Sinclair augmented with user experiments.**
- **Stubbe: Motivation. Genres organized in a hierarchy in order to reach a high coverage with respect to real world corpora and provide categories that are useful to support applications. Construction. Extension of a previous genre palette based on a user survey.**

3) Granularity of web documents

- **Puschmann: web feeds. There is a hierarchy of levels to how the authors looks at the degree of stylistic variation and other numeric indicators:**

  a) item (one post in a blog)

  b) source (all posts in a blog)

  c) subgenre (all blogs of a presupposed functional type, i.e. blogs written for marketing purposes)

  d) genre (all blogs in the entire corpus)

  Web feeds are the data source that the author uses, but technically the "smallest unit" in my corpus is the individual web log post.

- **Ringlstetter: individual web pages.**
- **Rosso: individual web pages.**
- **Rubleske:** individual web pages (including *any* pages (e.g., PDF, XML) that can be viewed using most late-version browsers.
- **Sharoff: individual web pages (compared against documents in BNC and RNC).**

- **Stubbe: individual web pages.**

4) Granularity and number of genre classes

- **Puschmann: One genre or subgenre (the corporate blog), depending on your definition. Since the authors use a functional definition of genre (a typified interaction with a shared set of communicative goals) he refers to corporate blogs as a genre and not a subgenre.**

The following functional subcategories exist inside that genre:

A. general/multipurpose blog

B. image blog

C. knowledge blog

D. product blog

E. small/medium business blog

F. strategy blog

This classification follows basic communicative interests that a company maintaining a blog might have:

- to build and solidify reputation (type B),

- to manage organizational knowledge (type C – this is oftentimes done internally, i.e. many of these blogs are not available on the Internet) ,

- to market products (type D),

- to publicize plans relevant to the overall corporate strategy (type F – these blogs are typically written by the senior management),

- Type A and E differ from this classification and therefore I only include them her for sake of completeness. Type A is a catchall category that applies to blogs that realize a broad variety of functions – in such cases, very often there is only one blog for the entire organization and usually the number of authors for that blog is basically limitless (see the Google example below). Type F is based on organization size and because I assume the communicative goals of small and medium-size businesses to differ from those of publicly traded companies, they are assigned this special category.

- **Ringlstetter: 3 genres at basic level: *blogs, catalogs (eshops), FAQs*.**
- **Rosso: 18 genres at basic level: (see Rosso's thesis p. 130)**
- **Rubleske:** 115 top-level genres! But as noted above, our objective was to maintain symmetry with the topical directory (Clusty.com) for experimental purposes, hence the wide and very shallow (only two levels) structure.
- **Sharoff: multidimensional classification, comprising five communicative intentions complemented with mode and audience parameters.**
- **Stubbe: 7 supergenres (i.e. top level classes; 32 genres (at the basic level) (see colloquium abstract).**

5) Similarity of genre classes or fuzzy genre labels

- **Puschmann: Not applicable now. The author is likely to replace the abovementioned scheme with an author-function matrix in the future. Such a matrix would allow for a range of authors and an range of functions that**

can be realized in a single blog, as oftentimes a company will maintain just one blog, with a large number of people from different departments posting on a variety of topics (cf. http://googleblog.blogspot.com/). The author would make such a classification fully dependent on the self-description of the stakeholders, i.e. he would ask them to classify their blog for me using the described matrix.

- **Ringlstetter: distant classes.**
- **Rosso: distant classes.**
- **Rubleske:** distant classes.
- **Sharoff: distant classes.**
- **Stubbe: distant classes.**

6) Genre classification scheme

- **Puschmann: SchemaCMD classification scheme (see Herring, 2007). More precisely, SchemaCMD for a technical classification of blogs overall, the abovementioned author-function classification for corporate blogs.**
- **Ringlstetter: one genre per web page.**
- **Rosso: one genre per web page.**
- **Rubleske** one genre per web page.
- **Sharoff: a tuple of labels per document.**
- **Stubbe: Single label and/or multiple labels per web page.**

7) Genre identification features

- **Puschmann: text descriptors (word and sentence length, POS frequencies, etc). In addition to the measures listed, the author uses Heylighen and Dewaele's f-score (cf. http://www.springerlink.com/content/p08225g588771321/), a measure based on part of speech distribution that captures context-dependency in texts. Texts with a high information density and explicit reference score significantly higher than texts that have frequent pronominal reference and depend on the reader's knowledge of the context.**
- **Ringlstetter: handcrafted feature sets (see Stubbe)**
- **Rosso: users' recognition**
- **Rubleske:** users' verbalized descriptions (of the page's 'type').
- **Sharoff: POS trigrams**
- **Stubbe: handcrafted features set (POS, wordlists, HTML, statistical features (average sentence length etc), compound features, …)**

8) Corpus format

- **Puschmann: XML. More specifically, Atom and RSS, which are both types of XML.**
- **Ringlstetter: HTML Files without pictures**
- **Rosso: Not Applicable.**
- **Rubleske:** html (text) and image files.

- **Sharoff: plain text.**
- **Stubbe: HTML Files without pictures**

9) Corpus storage

- **Puschmann: relational database (MySQL).**
- **Ringlstetter: File system, features stored in relational database (MySQL)**
- **Rosso: Not Applicable**
- **Rubleske:** relational database (MySQL) on an Apache server.
- **Sharoff: CWB, http://cwb.sf.net/**
- **Stubbe: File system**

10) Corpus construction and annotation:

- **Puschmann: Construction: web feeds as data source (web feeds are associated with RSS and Atom protocols). Annotation: Part-of-speech tagging is performed automatically via TreeTagger (cf.). Meta-data on the individual blogs (name of the company, gender of bloggers etc) has been added manually. Note that because of the granularity of the data there are only about 130 blog sources, but close to 23,000 posts from those sources that all share the same meta-data.**
- **Ringlstetter: Construction and Annotation: Re-use. Re-Use of 3 web genres of the 7-web-genre collection, built with the criteria of objective sources, & consistent genre granularity (see Santini, 2006).**
- **Rosso: Not Applicable**
- **Rubleske:** See #2 above. Corpus was constructed for experimental purposes. Genre taxonomy was built from users' verbal reports and researcher's informed judgments.
- **Sharoff: Construction: a snapshot of the Web comprising about 150 million words (see Sharoff, 2006), existing representative corpora (BNC and RNC). Annotation. Manual annotation of 200 documents for web corpora; use of existing classifications of the BNC and RNC.**
- **Stubbe: Constrution. Handpicked web pages (Annotation of random web-pages, using search engines to find specific genres). Annotation. Manual annotaion (one annotator, a small fraction checked by a second annotator)**

11) Size of the genre-annotated corpus and genre distribution

- **Puschmann:**

  ***All Corporate Blogs:**

  Blogs in this collection: 132

  Posts in this collection: 22,909

  Collection word count: 5,027,445

  ***general/mp Stats:**

  Blogs in this category: 24

Posts in this category: 2,092

Category word count: 596,304

***image**

Blogs in this category: 10

Posts in this category: 790

Category word count: 190,241

***knowledge**

Blogs in this category: 15

Posts in this category: 10,883

Category word count: 2,368,972

***product**

Blogs in this category: 31

Posts in this category: 3,966

Category word count: 678,757

***SMB**

Blogs in this category: 18

Posts in this category: 1,604

Category word count: 313,645

*** strategy**

Blogs in this category: 29

Posts in this category: 2,811

Category word count: 717,966

- **Ringlstetter: 600; 200 per 3 genre classes**
- **Rosso: Not Applicable.**
- **Rubleske**2,800 web pages – will provide update on genre distribution.
- **Sharoff: 200 web pages (in English); discussion=45%, information=11%, recommendation 34%, instruction=6%, recreation=4%. For Russian and German see Sharoff (2006)**
- **Stubbe: 1,200 web pages; 40 per 32 genre classes**

12) Corpus evaluation and assessment of corpus representativeness

- **Puschmann: There may be a certain degree of representativeness to to the fact that the corporate blog is a very young and narrowly defined genre, i.e. it is the author hopes that a substantial percentage of all corporate blogs currently in existence in indexed in the database. There is, however, no statistical measurement for this.**
- **Ringlstetter: No statistical evaluation; no representativeness assessment for the used corpora for positive examples. For the negative examples see above (Stubbe).**

- **Rosso: Not Applicable.**
- **Rubleske:** no statistical evaluation; no representativeness assessment (validation wasn't essential for our experiment).
- **Sharoff: application of confidence level measure to assess corpus representativeness (Sharoff, 2006)**
- **Stubbe: During construction it was taken care to have a wide distribution of topics, authors and webpages to ensure representativeness. No statistical evaluation.**

~~*~*