



Abstracts

Proceedings of the Colloquium *Towards a Reference Corpus of Web Genres*

Held in conjunction with Corpus Linguistics 2007

Birmingham, UK, - July 27, 2007

Organizers: Marina Santini and Serge Sharoff

COLLOQUIUM WEBSITE: [HTTP://CORPUS.LEEDS.AC.UK/SERGE/WEBGENRES/](http://corpus.leeds.ac.uk/serge/webgenres/)

CORPUS LINGUISTICS 2007 WEBSITE: [HTTP://WWW.CORPUS.BHAM.AC.UK/CONFERENCE2007](http://www.corpus.bham.ac.uk/conference2007)

Towards a Reference Corpus of Web Genres

Colloquium Schedule

27 July 2007

Lecture Room 2 (LR2)

First Part: 13:45 - 15:30

13:45-13:50	Welcome	
13:50-14:15	Serge Sharoff	<i>In the garden and in the jungle: comparing genres in the BNC and Internet</i>
14:15-14:40	Mark Rosso	<i>Development of a Genre Palette</i>
14:40-15:05	Joseph Rubleske	<i>Building a Corpus of Genre-Tagged Webpages for an Information-Access Experiment</i>
15:05-15:30	Andrea Stubbe	<i>Recognizing Genres</i>
15:30-15:45	Coffee Break	

Second Part: 15:45 - 17:30

15:45-16:10	Cornelius Puschmann	<i>SchemaCMD: An XML-based storage schema for the compilation of mixed-source CMD corpora</i>
16:10-16:35	Christoph Ringlstetter	<i>Incremental genre classification</i>
16:35-17:00	Rüdiger Gleim	<i>A Corpus Model of Structure Formation in Hypertext Types</i>
17:00-17:30	Final Discussion and Winding up	

Colloquium Description

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 (Lee, 2001) 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 2004, or TREC-2006 Blog Track), but the set of genres proposed was either debatable (e.g. the 'reaction' genre in TREC HARD 2003), or limited to a single genre (e.g. the blog genre in TREC-2006 Blog Track).

The web is new, so it is even less not clear how to apply traditional notions of genre to web documents. 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. Genre annotation has been based either on the common sense of a single rater, or on the 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 documents 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 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?

Rationale for the Colloquium

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 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. Kwasnik, 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.

Three further 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. 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. Finally, Cornelius Puschmann proposes an XML-based storage schema for the compilation of computer-mediated discourse (CMD) corpora from mixed sources.

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.

Programme Committee

Marco Baroni (University of Trento, Italy)
 Stefan Gries (University of California, USA)
 Adam Kilgarriff (Lexmasterclass, UK)
 Alexander Mehler (Bielefeld University, Germany)
 Sven Meyer zu Eissen (University of Weimar, Germany)
 Paul Rayson (UCREL, Lancaster University, UK)
 Georg Rehm (University of Tuebingen, Germany)
 Marina Santini (University of Brighton, UK)
 Serge Sharoff (University of Leeds, UK)
 Benno Stein (University of Weimar, Germany)

Organizing Committee

Marina Santini (University of Brighton, UK)

Email: MarinaSantini.MS@gmail.com

Personal Home Page: <http://www.nltg.brighton.ac.uk/home/Marina.Santini/>

Serge Sharoff (University of Leeds, UK)

Email: s.sharoff@leeds.ac.uk

Personal Home Page: <http://corpus.leeds.ac.uk/serge/>

Colloquium Abstracts

A Corpus Model of Structure Formation in Hypertext Types	7
Building a Corpus of Genre-Tagged Webpages for an Information-Access Experiment.....	8
In the Garden and in the Jungle: Comparing Genres in the BNC and Internet.....	10
Development of a Genre Palette	13
Recognizing Genres.....	14
Incremental Genre Classification.....	17
SchemaCMD: An XML-based storage schema for the compilation of mixed-source CMD corpora	22

A Corpus Model of Structure Formation in Hypertext Types

Alexander Mehler¹ and Rüdiger Gleim

¹Bielefeld University (Germany)

Short Abstract

This presentation 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.

Long Abstract

We argue in favor of a corpus model of structure formation in hypertext types or web genres, respectively. Our starting point is a layered graph model which distinguishes document networks, document units and their constitutive modules down to the level of elementary linguistic components. Consequently, we will distinguish web pages and websites as manifestation units of web genres. This model of the logical document structure of hypertext types focuses on the linkage of web genre constituents and, thus, investigates the distinction of traditional and web genres from a structural point of view. One advantage of this model is that it allows specifying the correlation of functional and structural varieties in web-based communication in terms of poly-functionality, polymorphism and vagueness. We will present a structural model which integrates this informational uncertain mapping and will describe a serialization thereof in terms of the Graph eXchange Language (GXL). Finally, we will provide a database mapping which allows retaining a wide range of web data for corpus linguistic and computational linguistic genre studies. This will be the starting point for a classification experiment which sheds light on Biber's structural hypothesis about the correlation of functional and structural varieties in the area of web mining. In order to do that we will exemplify three web genres, i.e. personal academic homepages, conference websites and city websites. Each of these genres will be mapped onto our graph model and we will provide statistical data which describes the distribution of document structures within these genres. A feature which makes classification of functional constituents of websites a very hard task is the Zipfian nature of the distribution of hyperlink-based document structures. In other words: the absolute majority of sites consist of single pages which hide, so to speak, the functional or thematic structuring of the site. At the same time we observe a rapid transition to more elaborate sites with kernel hierarchical structures superimposed by graph-inducing up, down, and across links. This Zipfian transition (which is mapped by a power law) is affected by the life cycle of the web genre in question: younger instances are more likely less structured than older ones. We argue that both observations have a high impact on web genre classification especially in those cases where the simple classification model of unstructured web genre tags is replaced by a model which is more sensitive to document structures.

Building a Corpus of Genre-Tagged Webpages for an Information-Access Experiment¹

Barbara H. Kwaśnik¹, Kevin Crowston¹, Joseph Rubleske¹ and You-Lee Chun¹

¹Syracuse University School of Information Studies (USA)

Short Abstract

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.

Long Abstract

This paper 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. Of interest to us at this part of our study is the presentation of results once a search has been carried out. That is, we're looking at the concluding end of the search-communication process where a user has communicated a need, and then must interpret the (usually) overwhelming set of results. Providing genre information might help in filtering the output, thus improving both efficiency and efficacy, as well as satisfaction with the experience.

Towards this end we have designed a set of experiments to test our premise, and have built a corpus of genre-tagged webpages to populate our test collection. In keeping with the colloquium's theme, our presentation is not about the experiment per se, but rather about building the corpus to be used in the experimental conditions. The identification of the genres was carried out as follows: We elicited names of genres from respondents from three domains (teaching, journalism, and education) who identified the genres of pages they visited while working on a real task for their own work. We recorded clues and labels, and roughly organized the genre terms into a shallow hierarchical taxonomy so that we could manipulate the granularity if needed.

Next, we used a clustering search engine to harvest potentially useful webpages for a set of 14 "canned" tasks/questions that we think will provide good opportunities for testing the effectiveness of genre information. An example of a task is to find a webpage that lists the countries that became independent on the same date as Guatemala. A clustering search engine (such as clusty.com) uses a proprietary algorithm to present search results in groups that can roughly be understood to be "topical." We say "roughly" because the clustering is in fact only approximately topical since it is based on keywords and at times the logic behind the choice of the keywords is not immediately apparent.

Once we harvested the roughly clustered pages (about 200 pages per task or 2800 in total), we tagged each page with a genre identifier using the previously devised genre taxonomy (from the field experiments) as well as supplemental terms as they were needed. The genre terms were chosen

¹ This research was partially supported by Grant 04-14482 from the US National Science Foundation.

by the page analyst. A small number of pages could not be coded because the analyst could not identify a genre.

In the experiments, which we have not yet conducted, the pages will be presented to the subjects clustered by keyword (as they were clustered by the search engine) and clustered by genre, and performance on the tasks in the two conditions will be compared (e.g., time, number of categories and pages examined, quality of solution and overall satisfaction).

In developing the corpus of webpages for the experiment, we have run into many issues, several of them so well articulated in the call for papers for this colloquium:

- It is often difficult for people to identify genres by name, even if they are clearly identifying a “genre.”
- Genres differ in their inclusiveness and specificity
- There are many equivalent or near-equivalent terms for the same webpage
- It’s difficult to unambiguously link a genre, to a task, or to the clues that were identified.
- Genre-identified webpages can be composed of pieces of other genres.
- Some reported “genres” are highly personal and idiosyncratic.

We report on the practical, working solutions to these questions for the purpose of creating our *experimental* (i.e., controlled) corpus. In other words, our labeling decisions for the purpose of this study were not intended to be a general-use genre taxonomy. While we had thought originally that we would be able to do so, in the process of developing a controlled experiment we found we had to find appropriate pages first, and label them second, not always following the genre terminology elicited from our informants. We imposed both labels and structure in order to make the two experimental conditions comparable. Thus, we aimed to make our genre terms understandable (which we’ll test in our pilot), but not necessarily 100% analogous to those we had originally collected in the field experiments.

We have structured this corpus in an admittedly somewhat artificial way so as to provide the maximum control for our experiments. At the same time we have tried to preserve the links to our “naturally” elicited genre terms in order to ground the experimental environment in what people actually do. In doing so, we recognize that much rich genre information was either too difficult to represent or had to be pared away. While our solutions were out of necessity pragmatically driven we feel they may still offer possible guidance for design principles in the future.

In the Garden and in the Jungle: Comparing Genres in the BNC and Internet

Serge Sharoff¹

¹University of Leeds (UK)

Short Abstract

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 (Kilgarriff and Grefenstette, 2003). In this presentation I will compare English and Russian Internet corpora against their human-collected counterparts 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).

Long Abstract

Unlike traditional representative corpora (e.g. the BNC), large corpora automatically collected from the Web (Joho and Sanderson, 2004, Sharoff, 2006) lack important information documenting them, such as their domains and genres. The task of classifying their texts and comparing their composition to traditional corpora is difficult for several reasons. First, no established classification of genres exists: practically every study uses its own list of genres, e.g. compare the 15 classes in the Brown Corpus to the 70 genres in David Lee's classification of the BNC to the 120 genre labels in the Russian National Corpus (RNC). Second, the relationship between traditional genres and genres existing on the Web is not clear. Third, we need reliable automatic methods for identifying genres of arbitrary webpages. The fourth problem concerns the very design of the genre inventory. If the goal is to classify every Webtext, the number of genres is too large to be listed in a flat list. Only within the genres of academic communication we can come across research articles (with different conventions applicable to the humanities, engineering or fundamental research in the natural sciences), as well as popular articles, reviews, books, calls for participation, emails, mailing list discussions, project proposals, progress reports, minutes of meetings, job descriptions, etc. The fifth problem concerns "emerging" genres: new technologies can offer new avenues for communication, which readily produce new genres, for instance, blogs, personal homepages or spam. To compare them to traditional sources we need a common denominator, such as communicative intentions underlying creation of texts in respective genres.

A starting point for establishing a set of communicative intentions can be taken from such studies as (Sinclair, 2003, Sharoff, 2004, Aires et al., 2005):

- **recreation** – such texts are written for leisure-time reading; the two important subclasses are fiction (science fiction, crime, etc) and nonfiction (biographies, memoirs, etc);
- **information** – such texts provide information about something and answer questions on what has happened, and how or why it happens; newswires and encyclopedic entries are typical examples of this text type;
- **instruction** – such texts explain how to do something; e.g. recipes, software man pages, etc, as well as more descriptive texts such as FAQs, tutorials and textbooks;

- **discussion** – such texts are aimed at discussing a state of affairs (including typical newspaper articles, academic papers, travel stories, etc); unlike purely informative texts they tend to present the opinions of their authors;
- **recommendation** – the purpose of such texts is to make you behave in a certain way; examples include propaganda and advertising.

In addition to this a study of genres should use parameters other than communicative intentions, e.g. the authorship (single or corporate), audience, publication medium, etc. Another potential class to this list is regulations, referring to laws, small print and similar. Experiments with users show that detection of these more abstract parameters is considerably more difficult than detection of simple genre labels by their look and feel. For instance, users know how a blog looks like, so if a page looks like a blog, it can be classified as a blog, whereas the choice between information and discussion, as the main aim of its creation is less obvious. However, if a list of genres includes a simple entry for blogs, it cannot be compared to anything in the BNC, whereas their function is similar to that of opinion columns in newspapers, and is different from them in the audience size, distribution mode and authorship. Also a list of news items with the possibility of leaving reader's comments looks very similar to blogs in its layout, but does not share the communicative function with them. This suggests the need to ensure the right balance between abstract parameters and look-and-feel features.

Not all genre labels can be mapped unambiguously, so this gives us 751 BNC texts to work with (2,542 texts were used for Russian, as RNC texts are considerably shorter). Then, we trained SVM classifiers on the frequency of POS trigrams describing individual texts, as well as the frequency of punctuation marks. As shown in (Santini, 2007) this is known to be the most reliable indication of genres, which is applicable to both traditional written texts and webpages. For English the procedure achieved 88% accuracy with 10-fold cross-validation, while being significantly lower (71%) for Russian, which can be explained by the free word order which makes POS trigram statistics sparser, especially on shorter texts of the RNC. Finally, we applied the models trained on the BNC and RNC to English and Russian texts from Internet corpora. The accuracy was tested on smaller samples of 200 webpages annotated manually, see (Sharoff, 2006). The accuracy drops (to 80% for English, 61% for Russian), but this still allows reasonable interpretation.

The results of the experiment show that recreational texts are seriously under-represented in Internet corpora: 27% in the BNC, 43% in the RNC, but only 4% in I-EN and 11% in I-RU (because of the larger number of pirated texts and exchanges of jokes on the Russian Internet). At the same time, the balance of other text types in the BNC and RNC is reflected in Internet corpora. The implication of these results for constructing a reference corpus of web genres is that they indicate the relative proportion of text types on the Internet and guide towards figures for balancing the reference corpus.

References

- Aires, R., Santos, D., and Alusio, S. (2005). "Yes, user! ": compiling a corpus according to what the user wants. In *Proc. Corpus Linguistics*.
- Joho, H. and Sanderson, M. (2004). The SPIRIT collection: an overview of a large web collection. *SIGIR Forum*, 38(2):57–61.
- Kilgarriff, A. and Grefenstette, G. (2003). Introduction to the special issue of the web as corpus. *Computational Linguistics*, 29(3):333–347.
- Santini, M. (2007). Automatic Identification of Genre in Web Pages. PhD thesis, University of Brighton.

- Sharoff, S. (2004). Towards basic categories for describing properties of texts in a corpus. In *Proceedings of the Forth Language Resources and Evaluation Conference, LREC 2004*, Lisbon.
- Sharoff, S. (2006). Creating general-purpose corpora using automated search engine queries. In Baroni, M. and Bernardini, S., editors, *WaCky! Working papers on the Web as Corpus*. Gedit, Bologna. <http://wackybook.sslmit.unibo.it>.
- Sinclair, J. (2003). Corpora for lexicography. In Sterkenberg, P. v., editor, *A Practical Guide to Lexicography*, pages 167–178. Benjamins, Amsterdam.

Development of a Genre Palette

Mark Rosso¹

¹School of Business

C. T. Willis Commerce Building
North Carolina Central University

Durham, NC 27707
919-530-6386

mrosso@nccu.edu

Short Abstract

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.

Long Abstract

This presentation describes a series of studies conducted for the purpose of building a genre palette to be used to improve the effectiveness of web searching. A major issue in palette development is the identification of what document categories should be used as genres. As genre can be defined as a “folk typology”, document categories must enjoy widespread recognition by their intended user groups (aka discourse communities), in order to qualify as genres. Three user studies were conducted to develop a genre palette and show that it is recognizable to users.

The first study was a survey of user terminology for web pages. Three participants separated 100 webpage printouts into stacks according to genre, assigning names and definitions to each genre. 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. Ten participants classified the same 100 webpages. A set of five principles for creating a genre palette from individuals’ sortings was developed, and the list of 48 was trimmed to 18 genres. The third study aimed to show that users would agree on the genres of webpages, when choosing from the genre palette. In an online experiment in which 257 participants categorized a new set of 55 pages using the 18 genres, on average, over 70% agreed on the genre of each page.

Difficulties of experimental design and future directions for the work will be discussed.

Recognizing Genres

Andrea Stubbe¹ and Christoph Ringlstetter²

¹CIS, Univ. of Munich (Germany)

²AICML, Univ. of Alberta, Edmonton (Canada)

Short Abstract

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.

Long Abstract

1 Genre Palette

We introduce a hierarchy of genres, based on the definition of genre in terms of form and function. Although other dimensions such as topic, authorship, or medium may influence the genre of a text, these are not regarded as part of the definition.

Our main goals were to reach high coverage with respect to real world corpora and to provide categories that are useful to support applications, as for example document retrieval. Among the challenges we had to meet were the following: choosing the right granularity of the hierarchy, selecting an operationalizable definition for each genre, and avoiding a meaningless miscellaneous category at the top level. In each case we integrated several leaf classes into a category of a first hierarchical level to assist problems in which a coarser scheme is more appropriate. Additionally, this allows hierarchical browsing and broadening/restricting of the initially chosen genre, when needed. One could think of a deeper hierarchy, but so far we have not done any experiments where a third layer would lead to better results.

Our hierarchy extends previous work by Dewe et al. (1998), using the feedback they received from a user study. Dewe et al. (1998) introduced eleven classes which sometimes did not adhere to our definition of genre: private and public homepages, for example, only differ in the addressed audience and thus have been merged into the category presentation.

The classes other continuous text and interactive pages, criticized as being too general, were split up. All evolving leaf genres were gathered into seven top level classes: Journalism, Literature, Information, Documentation, Directories, Communication and Nothing, a class for texts with no function or content. A first version of the hierarchy was refined by inserting a number of random files - a good method to detect missing classes. Table 1 shows our hierarchy.

A. Journalism	C. Information	D.3 Protocol
A.1 Commentary	C.1 Science Report	E. Dictionary
A.2 Review	C.2 Explanation	E.1 Person
A.3 Portrait	C.3 Receipt	E.2 Catalog
A.4 Marginal Note	C.4 FAQ	E.3 Ressources
A.5 Interview	C.5 Lexicon, Word List	E.4 Timeline
A.6 News	C.6 Bilingual Dictionary	F. Communcation
A.7 Feature Story	C.7 Presentation	F.1 Mail, Talk
A.8 Reportage	C.8 Statistics	F.2 Forum, Guestbook
B. Literature	C.9 Code	F.3 Blog
B.1 Poem	D.Documentation	F.4 Form
B.2 Prose	D.1 Law	G. Nothing
B.3 Drama	D.2 Official Report	G.1 Nothing

Table 1: A hierarchy of genres

2 Corpus Construction

For each genre we hand-collected 20 English webpages for training and 20 for testing, leading to a corpus with 1280 files. We choose to provide a first corpus for the complete spectrum of genres and hope to broaden the statistical basis by integrating material of other groups and collecting additional documents from the web.

We tried to gather a broad distribution of topics, authors, and websites for each class to avoid corpora biasing towards these features and to guarantee generalizability. Hardly more than two files in each class agree in any of these other features. That leads to a much greater effort than taking several examples from one website, but is necessary if the classifiers generated by these training files should be transferable to pages from other websites or subjects.

To facilitate good performance of the classifiers, the collection for the training corpora was restricted to prototypical documents. The documents were randomly collected and sorted into their categories while surfing the web. If not enough files could be found that way, search engines were employed using keywords we expected to occur in the specific genres. These keywords were precluded as features for the classifiers. It turned out that some genres are a lot more common (or easier to find) than others. The web-specific ones such as blogs, forms or online-shops/catalogues did occur very often, feature stories and bilingual dictionaries were especially hard to find.

3 Features and Classifiers

We created a set of hand-crafted classifiers, one for each genre. The construction of the classifiers is based on the fact that each genre is defined by specific features. We calculated the mean occurrence of candidate features within each class of the training corpus and by this decided whether they provide effective discrimination between the genres. If not, they were discarded. Although, at first glance, this method seems prone for overfitting, the risk is quite small as the features have been derived by linguistic knowledge and not by statistics.

4 Assigning Multiple Labels

One question which arises when talking about classification is, whether an item may fall into a single class, multiple classes, or sometimes even no class at all. As stated before, some texts genuinely belong to more than one class: epistolary novels are a mixture of letters and novels; blogs may contain several texts of different genres, such as poems or code listings, but still remain blogs.

These examples illustrate the two types of multi-class documents. The first one is a single text which simultaneously falls into several genres (or, to be precise, into a mixture of these genres), the second one is a collection of texts belonging to different genres. For this second type, a new genre collection might be introduced, defined by a contains-relation with the genres of the sub-texts.

Our approach acknowledges the need for assigning multiple labels to one document without distinguishing between the two types of natural multiple class documents, but also provides the possibility to restrict to one single label by "first come first serve" techniques.

5 Mono Classification

Two methods for choosing a single genre for each document were evaluated. For both, we introduced an ordering on the set of classifiers. A document is passed through an ordered sequence of classifiers and the processing stops as soon as the first classifier identifies the text as belonging to his class. The first approach arranges classifiers by F1 metrics, the highest first. Dependencies between the classifiers are not considered. A more sophisticated technique uses these interconnections to find a locally optimal sequence. The first version of the classification sequence is established by declining recall values, with precision as a secondary ordering criterion. We then use a dependency graph arising from the confusion matrix to rearrange classifiers: if a classifier (N_i) depends on a direct successor (N_j), that means N_i wrongly recognizes files belonging to N_j , the two classifiers change places. With this approach we diminish misclassifications and augment precision.

6 Experiments

When we applied the ordering arising from the dependency graph, we obtained the following results for the test collection. The precision of the classification into original classes was 72.2% with an overall recall of 54.0%. The quality of classification differed considerably between certain classes, ranging from an F1 value of 14.7% for *{\em marginal notes}* (A.4) to 100% for *{\em nothing}* (G.1). Genres with a definite gestalt such as directories, poems, FAQ, and forums were generally recognized above average. If we consider documents as correctly classified that do not end up in their original class but in a class that is also well-justified (such as a scientific report including a great part of statistical information that has been classified to statistics), the precision rises to 80.5%. Reducing the hierarchy to the more coarse grained first level, we obtained a precision of 77.8%.

7 Conclusions

The shortcoming of a small corpus is that the training of machine learning algorithms does not lead to satisfactory results, as these algorithms often require several hundred training examples, especially if - as in the case of genre - classes are fairly similar.

That made it necessary to spend more effort on crafting of the classifiers and selecting useful features. A great advantage of our corpus is that the documents have been collected from different sources, authors and topics. Thus, our classifiers work and generalize well, especially when regarding the humble size of the corpus. An additional strength is that all documents are carefully handpicked leading to a high quality of the training material.

References

Dewe, Johan; Karlgren, Jussi; Bretan, Ivan (1998). Assembling a Balanced Corpus from the Internet, In *11th Nordic Conference of Computational Linguistics*, Copenhagen, Denmark.

Incremental Genre Classification

Andrea Stubbe¹, Christoph Ringlstetter², Tong Zheng² and Randy Goebel²

¹ CIS, University of Munich (Germany)

² AICML, University of Alberta (Canada)

Short Abstract

Most approaches for the partition of document spaces into different genres rely on static training corpora. However, thinking of applications, for example within search engines, static classifiers disregard potentially valuable data available via explicit or implicit user feedback. We provide an initial scenario of incremental genre classification. A taxonomy of user behaviors is applied to develop possible strategies for classifier adaption driven by user feedback simulated using annotated corpus data.

Long Abstract

Genre as a selective dimension of an increasingly less concise document space is receiving more and more attention. An obvious application of genre classification is the refinement of document search. During the public employment of a genre interface, a steady stream of user events will arise. If these behavioral observations can be turned into meaningful data, they can be exploited to adapt the start configuration of the underlying classifiers.

1. Search interface

To give the user the possibility to restrict his document search to certain genres the usual search interface has to be adapted. A very simple explicit adaption is to augment the input window where the user enters his query by a *genretype attribute* analogous to the *filetype attribute* most of the current search engines provide. To enable an explicit feedback functionality, the result page has to be extended with a click box where the user can provide a statement on the genre of a presented webpage. Many variants of the sketched interface are conceivable with a completely *silent interface* as an extreme that is supposed to minimize the cognitive load of the user. Desired genres have then to be deduced from the gestalt of the query combined with locally or globally aggravated knowledge about the user. The feedback of the user with respect to the suggested genre labels has to be deduced from his observable navigation on the result set.

2. User behavior

To further analyze the proposed genre search interface we model four different scenarios of user behavior. We define a **query** as a non-empty set of keywords and a genre label. A **result set** is a set of ranked documents retrieved by the search engine processing a certain query. Each result document is annotated with a Boolean value referring to the genre selected by the user. According to our interface we define two different kinds of clicks: a **retrieval click**, the selection of a certain document, and an **evaluation click**, a user statement on the genre label of the document. The user's readiness to cooperate on the evaluation of the presented genre labels can be divided into four levels.

1.) **Fully cooperative behavior.** The user provides an evaluation statement for all documents of the result set: each page of the result set turns into correctly labeled data.

2.) **Cooperative behavior.** The user provides an evaluation statement of the annotation labels for the retrieved web pages. Thus, each retrieval click leads to an evaluation click.

3.) **Semicooperative behavior.** The user provides an evaluation statement only for a certain percentage of the visited pages.

4.) **Uncooperative behavior.** The user provides no information. Evaluation statistics can only be derived implicitly from the visiting statistics of the pages themselves.

According to studies of standard search engines, the average number of visited pages per search session is less than 2 and in most cases these 2 pages are retrieved from the first 20 hits of the search results. Consistent with this, we set, on average, a visit of two pages per turn. If both, labeled and unlabeled pages are present, the user visits the labeled pages. If the turn contains not enough labeled pages the user is assumed to be able to derive the desired genre with a certain accuracy from the snippet (*snippet recognition factor*). The resulting events are summarized in Table 1. For semi-cooperative behavior all are possible. Cooperative behavior is inconsistent with (iii) and (v). Uncooperative behavior excludes (i), (ii) and (iv).

(i)	user visits labeled page and confirms label
(ii)	user visits labeled page and rejects label
(iii)	user visits labeled page without evaluation
(iii.a)	page was correct classified
(iii.b)	page was false classified
(iv)	user visits unlabeled page and sets label
(v)	user visits unlabeled page without setting a label
(v.a)	page was correct negative
(v.b)	page was false negative

Table 1. A taxonomy of feedback events

3. Adaption of the genre classifiers

In previous work we have introduced specialized rule based classifiers that rely on aggressively pruned handcrafted feature sets. A necessary prerequisite to endow these static classifiers with the capability of adaptive response to new information is to rewrite them in disjunctive normal form (DNF). Generally, this implies each alternative rule combination to be linked to the other combinations by a logical **OR**. Within the disjunctive elements only connections by logical **AND** are allowed. Lower and upper bounds of the features' numerical ranges have to be explicit. Below we show a cut-out of the catalog-classifier in its DNF form.

$(\text{currency} > 3.1 \wedge \text{currency} < 100,000 \wedge \text{form} > 0.1 \wedge \text{form} < 100,000 \wedge \text{rel-curr.} > 1.51 \wedge \text{rel-curr.} < 100,000)$

∨

$(\text{currency} > 5.1 \wedge \text{currency} < 100,000 \wedge \text{form} > 0 \wedge \text{form} < 100,000 \wedge \text{rel-curr.} > 5.1 \wedge \text{rel-curr.} < 19.9)$

To establish comparability, all features are normalized to values within the interval [0..1]. The general adaption algorithm to process available information on the genre of an input file, given the premise of a static feature space, has to distinguish between two different situations:

a.) *False negative*: A document of genre N_i has not been recognized as N_i . For every disjunctive element of the classifier in DNF form, we compute the sum of the required range adaptations to achieve a correct classification of the input document. The element with the minimum sum is selected and its ranges are temporarily adapted.

Constraint: Generally, the files in the *relevant history* that are classified correctly attendant on the classifier adaption (*new correct positives*) have to outnumber the files that are now falsely classified (*new false positives*).

b.) *False positive*: A document of genre N_j has been falsely recognized as genre N_i . We identify elements of the disjunction that have approved the input document as N_i . Within the elements, we look for the smallest sum of adaptations that prevent the positive classification of the document.

Constraint: Again, the number of files for the *relevant history* that are classified correctly attendant on the classifier adaption (*new correct negatives*) has to be larger than the number of files that are now falsely classified (*new false negatives*).

Uncooperative user behavior: the challenge with uncooperative user behavior is to investigate the degree to which we can derive knowledge from events that do not involve explicit user statements. In practice and in literature the *lingering time* is used to substitute explicit relevancy judgments of a user. If the user stays at a retrieved webpage for a time longer than a certain threshold τ , the page is assumed to be relevant. The probability of the correctness of this assumption, $P(\text{relevant}(x)|\text{time}(x) > \tau)$, is estimated using relative frequencies within controlled user data. The inference from relevancy to the users' evaluation of genre labeling introduces additional difficulties. Either a correctly labeled page can be irrelevant for the user or an incorrectly labeled page can be relevant. To the best of our knowledge, the probabilities that judgments on the labeling derived by document relevancy are correct, $P(\text{genre}(x) = \text{label}(x)|\text{relevant}(x))$, $P(\text{genre}(x) \neq \text{label}(x)|\neg\text{relevant}(x))$, have so far not been investigated.

4. Experiments

In a first series of experiments on the incremental adaption of three example classifiers, *blog*, *catalog*, and *faq*, we used the corpus provided by Marina Santini for the positive examples, each split into 160 documents for training and 40 documents for measuring recall. For the training/test with negative examples we used a controlled corpus of 31 different genres. From the training corpora we randomly generated 48 result sets consisting of 20 documents, each containing ~ 3 documents of the desired genre, as the basis for the simulation of user behavior.

4.1. Experiments for the full cooperative user

The *full cooperative user* provides the interface with complete information about the binary classification of the presented data. In Table 2 we present the results for the adaption of the rule based classifiers and of an svm-classifier. For one genre, *faq*, the svm did not converge. Summarized, a significant improvement of the classification can be achieved by using fully labeled data. However, a fully cooperative user can only be expected if he has a very high personal interest in the improvement of the classification. To reconcile to a realistic search environment, we have to gradually adapt this concept.

Genre	Recall ^{Train}	Fallout ^{Train}	Recall ^{Test}	Fallout ^{Test}	Recall ^{Test-SVM}	Fallout ^{Test-SVM}
Blog	70.00 (61.25)	1.80 (0.50)	72.50 (57.50)	1.85 (0.13)	72.50 (65.00)	2.14 (1.07)
Catalog	58.75 (40.00)	1.32 (0.59)	52.50 (40.00)	1.19 (0.27)	47.50 (42.50)	1.37 (0.31)
FAQ	90.50 (41.50)	3.36 (1.33)	77.50 (52.50)	4.29 (1.20)	-	-

Table 2: Fully cooperative user. Results for adapted classifiers and start configuration (in brackets).

4.2. Experiments for the cooperative user

A rational cooperative user will retrieve pages of the desired genre and will give feedback whether they were correctly classified. If not enough positively labeled pages are available, it can be assumed that the user will try to derive the missing label from the snippets, retrieve the pages, and give feedback on the genre. As is immediately clear, the assumed user behavior of only retrieving two documents leads to a strong preference of events that can help to improve precision. The phenomenon of classification improvement despite of the data loss can be described as a case of *active learning* in that only a few interesting examples are sufficient to successfully adapt the borders of a classifier.

Genre	Recall ^{Train}	Fallout ^{Train}	Recall ^{Test}	Fallout ^{Test}	Recall ^{Test-SVM}	Fallout ^{Test-SVM}
Blog	81.25 (61.25)	6.40 (0.50)	83.40 (57.50)	6.36 (0.13)	72.50 (65.00)	2.14 (1.07)
Catalog	60.00(40.00)	1.73 (0.59)	52.50 (40.00)	1.06 (0.27)	45.00 (42.50)	1.98 (0.31)
FAQ	85.00 (41.50)	1.33 (1.33)	75.00 (52.50)	1.91 (1.20)	-	-

Table 3: Cooperative user. Results for adapted classifiers and start configuration (in brackets).

4.3 Experiments for the uncooperative User

With uncooperative user behavior, the lingering time of the user on a retrieved result page, depending on genre, topic and model exogenous factors, is transformed into a binary relevancy signal. A negative signal means that the document is irrelevant either because of wrong topic or because of wrong genre. Unfortunately, in a realistic scenario the topic precision is poor which prevents us from gathering reliable data on genre by a negative relevancy signal. This leaves the case where the lingering time exceeds the threshold. To get a positive relevancy signal for cases where the document is of the desired genre the topic must be relevant. Insofar, we have to expect data loss for correct positives and false negatives with a factor of $1 - \text{precision}(\text{topic}(x))$ and a small data gain via accidental confirmations by exogenous events. As for the documents of a genre different than that desired, we have false positives and correct negatives that can be amplified by a positive lingering signal caused by relevancy because of topic or by exogenous events. For our experiments we worked with deliberate probabilities of 0.1 for the lingering time caused by an exogenous event, 0.95 for a relevant document being of relevant topic *and* relevant genre, and a topic precision of 0.5. With these values we get a data loss of 45% for the correct positives and the false negatives and an defilement with 12% noise for the retrieved negatives. For the experiment with faq we received out of 48 result sets with 20 documents each, 0 feedback examples for false positives, 40 for correct positives, 6 for false negatives, 0 for correct negatives, 1 noisy example for correct positives and 7 noisy examples for false negatives. For both classifier types we get reduced but fairly robust improvements despite of the data loss and the defilement with noise.

Genre	Recall ^{Train}	Fallout ^{Train}	Recall ^{Test}	Fallout ^{Test}	Recall ^{Test-SVM}	Fallout ^{Test-SVM}
Blog	70.00 (61.25)	1.84 (0.50)	72.50 (57.50)	2.26 (0.13)	57.50 (65.00)	2.14 (0.15)
Catalog	56.25 (40.00)	1.22 (0.59)	52.50 (40.00)	0.97 (0.27)	45.00 (42.50)	0.92 (0.31)
FAQ	79.37 (41.50)	1.33 (1.33)	67.50 (52.50)	1.91 (1.91)	-	-

Table 4: Uncooperative user. Results for adapted classifiers and start configuration (in brackets).

5 .Conclusion

We have introduced an initial framework for the steady improvement of a genre search interface exploiting data of observed user events. Our next goals are to extend the simulation to more genres by collecting additional genre corpora and then to implement a prototype of a genre interface to collect real data for the estimation of now assumed values for the correlation between *lingering time* and correct genre and the *snippet recognition factor*. Finally, we want to extend the classifier adaption with respect to a dynamic feature space.

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

Cornelius Puschmann

Department of English Language and Linguistics
University of Düsseldorf, Germany

Short Abstract

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.

Long Abstract

While the Internet has been a treasure trove for linguistic investigation ever since its inception, the systematic collection of data specifically for linguistic purposes has been partly hampered by the fact that textual data existing on the Web in HTML format is not tagged semantically but visually and structurally. Because tag information in HTML documents provides rendering instructions to web browsing clients but includes very little (useful) meta-data, documents retrieved from the Web are not as contextually rich as they could be, often omitting information on the author, the exact time of creation, etc.

This paper presents a simple, web-based corpus-management tool working with a faceted structuring schema based on the eXtensible Markup Language (or XML) for the storage and linguistic analysis of Internet sources that use so-called data feeds (or web feeds, often associated with the RSS and Atom protocols) for syndication. As the breadth of user-generated data, for example in blogs, wikis and services associated with social networking sites continually increases, such feeds are posited to become a standard method of content distribution - a channel that retains the meta-information that was previously omitted.

The specific investigative focus of SchemaCMD will be to highlight speaker variation as a significant factor in corpus linguistic analysis. Because content from web feeds can virtually always be directly linked to its respective author by merit of the feed meta data, such texts allow for a rich contrastive analysis of individual language production. This makes the compilation of large corpora that capture variation over time, variation between different texts produced by the same author and variation between texts written by different authors possible, which adds a new dimension to corpus linguistics in the area of computer-mediated discourse.

An example: the corporate web log corpus (CBC)

For my thesis work on The Corporate Blog as an Emerging Genre of Computer-Mediated Communication I decided to build a representative corpus of company blogs. The above-mentioned corpus tool that integrates the SchemaCMD classificatory scheme (following Herring, 2007) was built for this purpose, though it can be used for any corpus that relies on web feeds as a data source. The fact that feeds form the basis of the corpus makes an in-depth contrastive and diachronic

investigation of corporate blogs as a text type possible. As data taken from web feeds is already represented in XML, the storage of such information in a relational database (MySQL for my corpus) was merely a matter of processing and copying the data. Below is a typical feed entry, in this case taken from McDonald's Open for Discussion blog. The Magpie RSS library for PHP is used to fetch the entry as an array:

```
Array
```

```
(  
  [title] => A Visit to McDonald's  
  [link] => http://csr.blogs.mcdonalds.com/default.asp?item=256774  
  [description] =>
```

```
Last week, Julia Hailes, a co-founder of SustainAbility, joined us in Chicago at  
our Corporate Relations Conference.
```

```
Julia participated in a panel discussion [...]
```

```
[comments] => http://csr.blogs.mcdonalds.com/default.asp?item=256774  
[pubdate] => Wed, 02 May 2007 10:11:15 EDT  
[author] => undisclosed@blogs.mcdonalds.com (csr)  
[guid] => http://csr.blogs.mcdonalds.com/default.asp?item=256774  
)
```

There are a number of issues with and differences between different version of the RSS and Atom specifications that need to be considered before content from a feed can be retrieved. For example, the fields description (RSS) and summary (Atom) can both be used to include the full text of an entry or merely the first n words (50 words is one typical cut-off point). Feeds using the atom Atom 1.0 specifications usually encode the full text of entries in the appropriate content field instead, reserving the summary field for a summary in addition to the complete entry. Other possible sources of errors include date format conversion from RSS/Atom to MySQL and the large number of fields in both protocols which are optional, as they can't be relied on in all further processing steps. The data structure of the posts table inside the MySQL database thus closely mirrors the structure of the feed.

Table structure for cdb.posts:

```
post_id  
post_blog  
post_title  
post_link  
post_fulltext  
post_author  
post_category  
post_comments  
post_guid  
post_pubdate  
post_stats_wc  
post_stats_tc  
post_stats_sc  
post_statsawl  
post_statsasl  
post_wordlisted  
post_processed  
post_omitted  
post_realdte
```

The fields title, link, fulltext, author, category, comments, guid and pubdate are taken directly from the feed, while the other fields store statistical data and internal flags to indicate which processing steps have already been performed and which are still to follow. All entries in the posts table are relationally tied to their respective parent items in the blogs table via ID numbers.

TABLE	RECORDS	DESCRIPTION
blogs	144	blog data
bnc	100	BNC list of 100 most frequent English words (for comparison)
categories	6	Manually created labels that can be applied to blogs (i.e. "product blog", "PR blog" etc)
collections	6	Manually created labels that can be used to created sub-corpora (e.g. "blogs", "press releases" etc)
gram2	165,789	2-grams by frequency(*)
gram3	280,338	3-grams by frequency(*)
gram4	318,056	4-grams by frequency(*)
gram5	325,720	5-grams by frequency(*)
pos	23,210	parts of speech frequencies by post
posts	23,210	post data
poststats	23,210	extended post statistics
tokens	5,920,583	tokens (1 row for each single word in the corpus; dependent on the types table)
types	210,393	types (unique strings) in the corpus; referenced by the tokens table
13 table(s)	7,290,765 (244.1 MiB)	
(04 May 2007)		
* while a function for counting n-grams (and pos-grams, because of the structure of the types table) has been implemented it is not currently in use due to the high computational cost associated with performing the necessary calculations for each entry		

When the blogs that are being tracked are checked for new content, the feed data provided by MagpieRSS is compared with the record inside the posts table. If the feed contains entries with guid values which do not occur in the table, it is assumed that they are new and they are copied to MySQL. Duplication of old items is avoided by checking guid values, or, in those cases where guid is omitted in the feed, by checking title – pubdate combinations, as they can also be assumed to be unique.

Other tables are used to store derived statistical data and information that is relevant for the internal function of the tool. The following overview lists each table in MySQL, along with the current number of records and a description.

Once all new entries have been recorded, the individual items are wordlisted using [TreeTagger](#). To do this as simply as possible, the main text of a post is written to a plain text file buffer (named `post.txt`) on the hard drive. TreeTagger is then called with the options `-token -no-unknown` and `-quiet` and the result of the tagging process is written to another buffer (`post-tagged.txt`). Finally, the tagged text is further processed with PHP and the result is written to the types and tokens tables, creating one row in the tokens table for every word and a new record in the types table if the string in question has not been previously recorded.

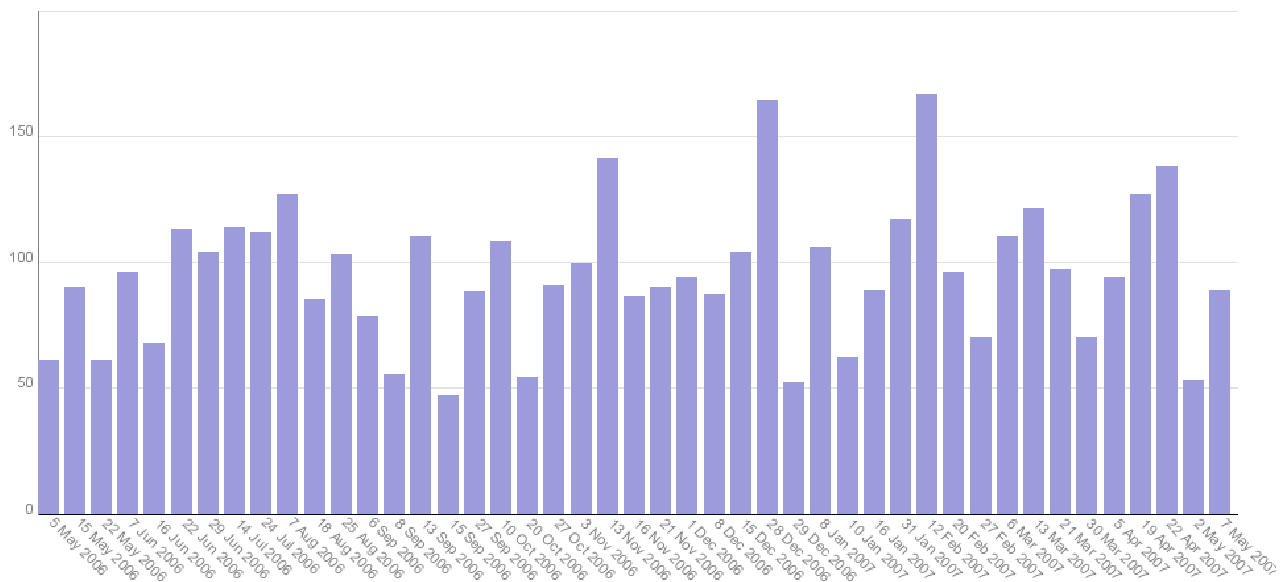
In a third step, extended statistical data is then computed and written to several different tables in the database. Virtually all of this information could also be computed on the fly, whenever the researcher needs it (e.g. the average word length for a given blog post), but it makes much more sense in terms of computational cost to make these calculations once and store and retrieve them as needed, especially in the case of cumulative computations where several results depend on one another. The following values are computed either upon indexing or later, when the researcher evaluates the data.

- word (token) count (WC)
- type count (TC)
- sentence count (SC)
- average word length (AWL)
- average sentence length (ASL)
- part-of-speech frequencies (POS chart)
- word frequencies (wordlist)

An interesting dimension in this context is that this data can be compared on multiple levels, i.e. for a single post, for all posts in a blog and for multiple blogs grouped together using a range of criteria. The extreme uniformity in how blog entries are segmented makes these comparisons feasible and allows for a detailed contrastive evaluation of the data.

The f-score (Heylighen & Dewaele, 2001) is a basic formality measure that is computed via the following frequency calculation: $0.5 * ((N + ADJ + PRP + DET) - (PN + V + ADV + ITJ) + 100)$.

The following table shows f-score variation over time for the abovementioned Open for Discussion blog.



This and other scalar measures can be used to track and evaluate the text production over time for a range of sources, which can be in turn compared contrastively inside the boundaries of an (assumed) functional genre. The advantage of such an approach is its greatly improved granularity:

variation is no longer assessed *only* between genres, but inside of them and among individual datapoints as they are provided by a single source.

Implementing an annotation scheme such as SchemaCMD thus ultimately serves the purpose of creating a computationally exploitable link between a text and its producer that takes as many contextual variables as possible into account.

References

Herring, S.C. (2007). A Faceted Classification Scheme for Computer-Mediated Discourse. *Language@Internet*. <http://www.languageatinternet.de/articles/761>. Retrieved 2.5.2007.

Heylighen, F. and Dewaele, J.-M. (2002). Variation in the contextuality of language: An empirical measure. *Foundations of Science*, 7, 293-340.