Towards a Reference Corpus of Web Genres for the Evaluation of Genre Identification Systems

Georg Rehm[1], Marina Santini[2], Alexander Mehler[3], Pavel Braslavski[4], Rüdiger Gleim[3], Andrea Stubbe[5], Svetlana Symonenko[6], Mirko Tavosanis[7], Vedrana Vidulin[8]

University of Tübingen, Germany[1]
SFB 441: Linguistic Data Structures

DSV, Sweden[2]
KTH-Stockholm University

University of Bielefeld, Germany[3]
Computational Linguistics Dept.

Inst. of Engineering Science, RAS[4]
Ekaterinburg, Russia

conject AG[5]
Munich, Germany

Nitol, LLC[6]
Moscow, Russia

Università di Pisa, Italy[7]
Dipartimento di Studi italianistici

Jožef Stefan Institute[8]
Ljubljana, Slovenia

Corresponding author:
georg.rehm@uni-tuebingen.de

Abstract

We present initial results from an international and multi-disciplinary research collaboration that aims at the construction of a reference corpus of web genres. The primary application scenario for which we plan to build this resource is the automatic identification of web genres. Web genres are rather difficult to capture and to describe in their entirety, but we plan for the finished reference corpus to contain multi-level tags of the respective genre or genres a web document or a website instantiates. As the construction of such a corpus is by no means a trivial task, we discuss several alternatives that are, for the time being, mostly based on existing collections. Furthermore, we discuss a shared set of genre categories and a multi-purpose tool as two additional prerequisites for a reference corpus of web genres.

1. Introduction

The field of automatic web genre identification is still in its infancy as an established research area.[1] Current approaches can be characterised as being highly heterogeneous: they usually work on a collection of web documents compiled by the researchers themselves; a category set is constructed and applied, so that all documents are tagged with one or more genres contained in the category set; finally, genre categorisation experiments are carried out.

Due to the success of widely used collections such as Reuters-21578, the Enron mail corpus, or the Penn Treebank in other application and evaluation scenarios it is obvious that there are severe problems inherent to isolated approaches as sketched in the previous paragraph. In other words, a reference corpus and a shared category set are needed. Currently, there is no such genre benchmark corpus against which to measure the performance of genre identification systems. Only a common dataset can enable researchers to compare and to evaluate their systems and to discuss interoperability issues. A reference corpus could prevent people from investing large amounts of time and money to come up with proprietary solutions to the complex tasks of building a corpus and a suitable category set.

First, section 2 discusses additional reasons for planning and constructing a reference corpus of web genres. Section 3 gives a short introduction into the field of web genre identification. We conducted an experiment among the authors in order to see how well experts in genre-related research perform in the task of assigning genre labels to web documents (section 4). Next, section 5 looks at the most important prerequisites for a reference corpus. These are shared sets of categories (section 5.1), collections of web documents (section 5.2), and a tool for the annotation of the collection with specific categories so that a gold standard benchmark can be built (section 5.3).

2. Rationale

A classification by genre can be of great value in a wide range of disciplines. For instance, in Information Retrieval, the concept of genre could help filter out irrelevant documents returned by keywords. Currently, keywords mostly express the *topic* of a document (for example, *politics*, *sports*, *football*, *international affairs*, *finance*), i. e., what a text is about, while *genre* expresses, in very general terms, the type of text (for example, *newspaper article*, *technical report*, *PhD thesis*, *scientific article*, *weather report*). In Information Extraction there are several approaches to identifying and extracting useful and relevant content from web pages, (Gupta et al., 2006) used a classification of websites based on genre and layout for this purpose. In Information Science, automatic genre classification could be useful for the automatic extraction of metadata essential to the efficient management and use of digital documents, especially in digital libraries. In Natural Language Processing, parsing accuracy could be increased if parsers were tested on texts that belong to different genres, as certain constructions may occur only in certain types of texts. The same is true for part-of-speech tagging, word sense disambiguation and related applications. More accurate NLP tools could in turn be helpful for the task of automatic genre identification, because many features used for this task are extracted from the output of taggers and parsers, such as POS frequencies and syntactic constructions. In Corpus Linguistics, automatic genre classification could help in the construction of diversified and more balanced corpora.

The availability of a reference corpus of web genres – web documents that have been tagged with genre and web genre labels – would be advantageous for all of the fields and application scenarios mentioned above. It could provide

[1] This article is the result of a discussion the authors had at the workshop "Towards Genre-Enabled Search Engines: The Impact of NLP", held, in conjunction with RANLP 2007, on September 30, 2007 (Rehm and Santini, 2007).

a common ground for genre-related research and scientific discourse and it could be used as training data for machine learning approaches for tasks such as automatic web genre identification. Furthermore, such a reference corpus would be a very interesting object of research for Linguistics, and applied work in Computer Science, Information Architecture, and Web Design. Linguists, especially text linguists, study, among other aspects, how genres form and develop, what their constituent elements are and how they can be described in the most adequate way. Related research questions are concerned with identifying culture-specific as well as universal aspects of genres. A multilingual reference corpus could provide a shared resource for comparative studies of this kind. In Computer Science, a reference corpus could be used as a resource to improve crawling algorithms, spam detection (this includes the detection of link farms) and web mining. Finally, a reference corpus could be a valuable tool for web designers and information architects who would have at hand a reference tool that includes example documents for multiple web genres that could be used as typical documents (both current and historical) or best-practice blueprints for new documents, see also (Ivory and Hearst, 2002) and (Rehm, 2007).

3. Automatic Web Genre Identification

While keywords express the *topic* of a text, *genre* expresses its type. Keywords can be ambiguous, even misleading – this is why keyword-based searches frequently return irrelevant results. The concept of genre helps in distinguishing different types of texts, e. g., *academic paper*, *manual*, *editorial*, and *blog*. These genres show characteristics that are – mostly – topic-independent. In an IR system, genre and topic should be, ideally, used together to increase its accuracy, so that queries such as "*academic papers* about *global warming*" could filter out texts of other genres.

Preliminary results in genre-enabled IR were reported by (Karlgren et al., 1998). (Xu et al., 2007), (Yeung et al., 2007) and nearly all other approaches since the seminal papers by (Karlgren and Cutting, 1994) and (Kessler et al., 1997) suffer from the same shortcomings: genre category sets are built according to subjective criteria for corpus composition, genre annotation, and genre granularity. The field is characterised by small-scale, self-contained, and corpus-dependent experiments. The lack of a reference corpus of web genres makes it impossible to compare these experiments and to evaluate progress. For instance, it is more or less impossible to compare the genre classification results reported by the studies listed in table 1. Is the 92% accuracy achieved by (Boese, 2005) better than the 70% accuracy obtained by (Meyer zu Eissen and Stein, 2004)? As table 1 illustrates, the following variables differ in approaches: corpus size, number of annotators, number of genres, number of web pages per genre. Furthermore, studies usually do not make explicit their annotation criteria or the composition of their category sets of web genres.

Genres can be analysed at various level of granularity. Some researches focused on super-genres and thematic classes (Vidulin et al., 2007), or created hierarchies of genres (Stubbe and Ringlstetter, 2007). Others yet used the functional styles belonging to the Russian linguistic tradition (Braslavski, 2007), or used functional classes derived from the lexicographic tradition (Sharoff, 2007b).

While the authors cited so far created their collections using the individual web page as the main unit of analysis, another line of research proposes to investigate genres at the level of websites. (Symonenko, 2007), for instance, identifies genre-like regularities in the structure of commercial and educational websites; (Björneborn, 2008) examines nine institutional and eight personal meta-genres in university websites; (Littig and Lindemann, 2008) present an approach for the automatic classification of websites into eight super-genres by combining content and structure. (Mehler, 2008) emphasises the discriminating power of structural information and applies his approach to different types of complex documents, from German thematic classes, to web genres at the level of websites (*city website*, *conference website*, and *personal academic home page*), to complex networks, such as, for example, wikis. While some researchers focused on automatic classification, e. g., (Kennedy and Shepherd, 2005) or the analysis of a single genre (Tavosanis, 2007), others built corpora such as the one by (Kim and Ross, 2007a) that includes 70 genres identified in a large collection of PDF documents; finally, several researchers are interested in fine-grained genres, e. g., (Rehm, 2007), (Levering et al., 2008).

Recently it has become popular to test classification approaches over several existing web genre collections. This cross-testing technique has been adopted by (Dong et al., 2008), (Kim and Ross, 2007b), and others. On the way towards a more objective evaluation of classification results, this technique can be considered a significant step forward, but it only partially addresses the issues underlying the need for a more objective assessment of genre classification approaches, because existing genre collections have been built without the ambition of being genre reference corpora. Consequently they do not fulfil the requirements usually associated with a reference resource.

For several reasons the construction of a genre reference corpus is an extremely difficult endeavour. One of the most important problems concerns the elusiveness of the concept of genre. The consequence is that, in practical terms, genre researchers usually have different ideas of what a genre is, how genres should be defined and identified and, therefore, they use different genre labels in their approaches – this is particularly evident in the experiment reported in section 4. The situation does not improve when non-experts, and presumably non-prejudiced annotators are asked to label web pages by genre (Rosso, 2008; Santini, 2008).

The proliferation of genre classes cited in the literature varies in terms of generality (super-genres, genres, subgenres, functional styles, functional classes, relatively arbitrary text classes or groups of texts that share a certain property, the odd topical category as well as "misc" or "other" classes). They are more or less influenced by domain-specific, linguistic, or structural features, and analysed at different levels of document granularity (page segment, web page, website, network). How can we convey this variety in a reference corpus that, although designed to be large, is necessarily limited in size? Additionally, as we do not know the distribution of genres on the web, we

Authors	# web pages	Annotation	Labels	Accuracy
(Santini, 2006)	1400	Annotation by the criterion of "objective sources"	blogs, eshops, FAQs, front pages, listings, personal home pages, search pages	ca. 90%
(Finn and Kushmerick, 2006)	2150	single rater	Subjective vs. objective and positive vs. negative	ca. 79%/49%
(Boese, 2005)	343	The author plus at least one or more raters	abstract, call for papers, FAQs, hub/sitemap, job description, resume/C. V., statistics, syllabus, technical paper	ca. 90%
(Kennedy and Shepherd, 2005)	321	n. a.	home page genres (personal, corporate, organisational)	ca. 70%
(Lim et al., 2005a, 2005b)	1224	two graduate students	personal home page, public home page, commercial home page, bulletin collection, link collection, image collection, simple table/lists, input pages, journalistic material, research report, official materials, FAQs, discussions, product specification, informal texts (poem, fiction, etc.)	ca. 75%
(Meyer zu Eissen and Stein, 2004)	800	three people, one of the authors checked some pages	article, discussion, shop, portrayal (non-private), portrayal (private), link collection, download	ca. 70%
(Lee and Myaeng, 2004)	321	at least two raters	reportage-editorial, research article, review, home page, Q&A, specification	ca. 90%

Table 1: A few recent studies in automatic web genre identification

cannot make any assumption about the proportion of each genre in the corpus. All genres are interesting in one way or another, but under the current conditions, it is difficult to claim that a genre reference corpus is representative.

4. Assigning Genre Labels to Web Pages: A Preliminary Case Study

The construction of a reference corpus necessarily involves the task of assigning linguistically motivated labels to documents by a group of annotators – in our case, the labels correspond to the names of genres or web genres. We conducted an experiment in which we wanted to measure the agreement of labels assigned to a random sample of 50 web documents by persons who are engaged in genre-related research.[2] Seven of the nine authors participated in this experiment. Other studies have shown that user-based genre labeling usually exhibits a certain kind of fragmentation and a low, at most moderate, inter-rater agreement (Rosso, 2005; Santini, 2008). In other words, the assignment of genre labels to documents, especially web documents, is a very hard task. The aim of our experiment was to see whether genre expertise is able to improve inter-rater agreement. The result we expected was that the labels assigned by the participants make some kind of sense, that probably about half of the genre categories are synonymous, or at least very similar and, if they are not, that they can at least be reduced to the same basic concept or text type.

The URLs of 50 randomly selected web pages were set up on a wiki page, and the seven participants noted their genre labels in a spreadsheet. There were no predefined annotation criteria or guidelines for assigning genre labels; multi-labeling was allowed. Researchers annotated the web pages without knowing the labels assigned by other participants in order to avoid bias.

The genre categories assigned by the seven experts contain a high number of disparate labels, ranging from genres and super-genres, to descriptions, to functional or purpose-oriented properties of a document, to topical categories. For instance, the document shown in figure 1 was labelled as

[2]The list of documents and labels is available in our wiki system, see `http://129.70.40.20/WebGenreWiki/index.php5`.

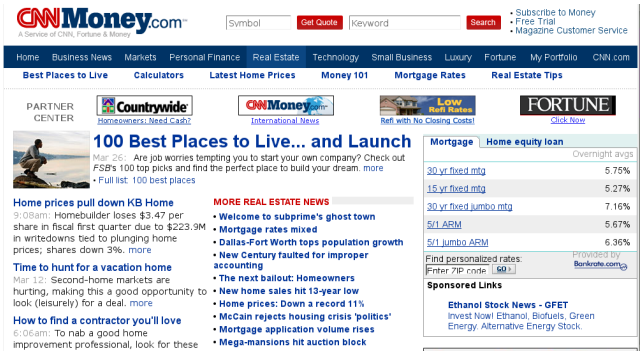

Figure 1: One of the 50 documents used in the experiment

"homepage", "journalistic", "department entry page of a news website", "a topic-specific section of an information portal", "portal", and "composite informational".

Consistency	No. of annotators who used same genre label	No. of documents	
High	5 to 7	6	12%
Medium	3 or 4	26	52%
Low	1 or 2	18	36%

Table 2: Results of the label assignment experiment

Due to the highly heterogeneous results it was impossible to perform any statistical analysis or to measure inter-rater agreement. Therefore, we assessed the annotation by emphasising the consistency of label assignments (see table 2). The qualitative analysis of the genre assignments clearly shows that the majority of genre labels are not consistent – most surprisingly, even researchers with a certain amount of genre-related expertise cannot be expected to come up with similar genre labels. However, a certain level of agreement exists for familiar genres with very large discourse communities, for example, *blog*, *academic article*, and *newspaper article*. Another phenomenon concerns the level of abstraction or generalisation applied by the participants to come up with a genre label: an article on a specific type of car with a novel hybrid engine was tagged as "article" (twice), "review" (twice), "advertisement/reportage", "a new product infomercial",

and "journalistic". Similarly, one of the patents available on http://www.freepatentsonline.com was labeled as "patent" (twice), "patent specification", "a patent page", "law", "scientific", and "[table-of-contents] with snippets". Moreover, the impact or importance of a specific website is able to overshadow or to influence genre assignment: while some participants used categories such as "entry in an online encyclopedia", "encyclopedia entry", or "encyclopedia" to label an entry in the Wikipedia encyclopedia, two other participants used "wikipedia entry". Most interesting are the labels assigned to documents 12 (http://pra.aps.org) and 13 (http://pubs.acs.org/journals/jpcafh/). Though immediately adjacent in the sample, only three participants realised that both documents have identical genres. They tagged the documents as "homepage", "homepage of a subscription-based academic journal", and "entry page of the website of a research journal" respectively. The other four participants assigned the inconsistent and diverse labels "composite informational", "newspaper, portal", "about-page", and "commercial/promotional" (document 12), as well as "table of contents with snippets", "portal, link collection", "bibliography/list of articles", "index, content delivery" (document 13).

From this small-scale experiment we can draw two conclusions. The task of assigning genre labels to web documents is, even for linguists, very hard. What is needed to arrive at a consistent set of genre labels are annotation guidelines that provide, in a detailed, transparent, and unambiguous way, a set of ground rules that explain the task of assigning genre labels to web documents.

5. Towards a Reference Corpus of Web Genres: Three Prerequisites

There are three essential prerequisites for a reference corpus of web genres. We need a shared category set that the majority of researchers in this field agree upon (section 5.1), a document collection (section 5.2), and an annotation and processing tool (section 5.3).

5.1. One or More Shared Category Sets

Before we can construct a reference corpus of web genres, the majority of researchers have to agree upon a shared annotation or categorisation approach that should be as precise and unambiguous as possible and, in addition, it should be possible to operationalise the approach. Furthermore, we need one or more shared category sets, because different or incompatible genre category sets make comparisons and evaluations impossible (see tables 1 and 3). Several category sets do not contain proper genres or web genres but categories that are topical or functional in nature (again, see table 3); in other words, some category sets are not based on the established terms, concepts, and distinctions used in textlinguistics and genre theory, but they contain categories that have been created in an ad hoc fashion (e. g., *discussion*, *simple tables/lists*, *person*, *resources*, *childrens'*, *subjective*, *official materials*, *content delivery*, *informative*).

A very crucial aspect concerns the scalability of current approaches. Compared to the number of genres or text types (*Textsorten* in the German textlinguistics tradition) identified by linguists, current studies on the automatic identification of genres only use very limited category sets. (Dimter, 1981) collected a list of more than 500 different genre labels from a German dictionary, while a count by (Ferrari and Manzotti, 2002) places the number of Textsorten described or referred to in a large commented bibliography (Adamzik, 1995) at "more than 4,500".

Another major obstacle is that some approaches assume that web genres can be adequately categorised on the "page" or "document" level alone. In addition to the assignment of genre categories to complete HTML documents, genres also work on an intra-document, or page segment level because a single document can contain instances of multiple genres, e. g., *contact information*, *list of publications*, *C. V.*, see (Rehm, 2002; Rehm, 2007; Mehler et al., 2007). In addition to a second category set for the web genre modules/components that occur on the intra-document level, we need a third category set, because web genres can be instantiated on the level of whole websites (Mehler and Gleim, 2006; Symonenko, 2007). Ideally, conventionalised connections between these three levels should be represented within the category sets (for example, that a *conference website* contains, among others, a *call for papers*, and a *schedule*).

There are several options we can pursue to arrive at a shared set of genre categories. We can adopt a bottom-up approach and derive the set from a document collection that has a very broad scope (option A). Alternatively, we can construct the set by exploiting the knowledge of researchers who work on genres and not taking into account any actual documents, thus, following the top-down method (option B). Both options have advantages and disadvantages.

Particularly, option A would result in yet another category set, adding to a heterogeneous collection of already existing genre sets developed by various researchers. Thus, it is very unlikely to facilitate the interoperability of the shared resource that we want to construct. In addition, the diversity of the texts and documents available on the web remains staggering and is changing at a rapid pace. Therefore, no matter how large its document base is going to be, the reference corpus will never be broad enough to cover all or even the majority of genres in existence online. A genre set developed following option A would necessarily suffer from the same problems as the genre sets created earlier: it would not be representative enough and it would be unable to support the purpose of a shared resource.

Option B (using expert knowledge to build a set), however, is also problematic. First, given the recency and the dynamics of web genres, it would be very difficult, probably unfeasible, to get the expertise of the required breadth and depth (Brandl, 2002). This option is likely to yield a less representative and more biased set of genre categories than option A. Moreover, without an actual document collection, there would be no valid reference for genre categories.

There is a third way of taking advantage of accumulated expert knowledge, which appears to be the most promising: a set of genre categories can be derived from collecting the category sets suggested by various research groups (option C). This task is by no means just a simple compilation of labels, but involves the evaluation, and refinement of categories in order to arrive at a consensus on the

(Meyer zu Eissen and Stein, 2004)	Help; Article; Discussion; Shop; Portrayal (non-private); Portrayal (private); Link Collection; Download
(Lim et al., 2005)	Personal homepages; Public homepages; Commercial homepages; Bulletin collections; Link collections; Image collections; Simple tables/lists; Input pages; Journalistic materials; Research reports; Official materials; Informative materials; FAQs; Discussions; Product specifications; Others
(Stubbe et al., 2007a)	Journalism (Commentary; Review; Portrait; Marginal Note; Interview; News; Feature Story; Reportage); Literature (Poem; Prose; Drama); Information (Science Report; Explanation; Recipe; FAQ; Lexicon; Word List; Bilingual Dictionary; Presentation; Statistics; Code); Documentation (Law; Official Report; Protocol); Directory (Person; Catalog; Resources; Timeline); Communcation (Mail/Talk; Forum; Blog; Form); Nothing
(Vidulin et al., 2007)	Pornographic; Blog; Childrens'; Commercial/Promotional; Community; Content Delivery; Entertainment; Error Message; FAQ; Gateway; Index; Informative; Journalistic; Official; Personal; Poetry; Prose Fiction; Scientific; Shopping; User Input
(Braslavski, 2007)	Official, academic, journalistic, literary, and everyday communication style

Table 3: Several recent genre category sets

category set's structure and the meaning of individual categories. In this way, we can avoid duplicating past work and put more effort into improving the resulting category set. Plus, the genre category sets we use have been created based on actual document collections. These datasets are already available and will form a core collection that we plan to expand into a reference corpus of web genres. Thus, option C appears to be most practical option and also the one that fits our ultimate goal of combining the research effort and building upon prior research. Table 4 shows the current state of our discussion (which is still ongoing). This list of genre categories is mainly geared to be applied to the level of complete HTML documents.

Among the group of authors of this article we discussed several options with regard to the nature of the categories. We do recognise that, in addition to genuine genres, other textual categories exist that operate on similar levels and that can be mistaken for genres at first glance. Among these are, for example, functional styles such as "official style", "academic style", or "journalistic style" (Braslavski, 2007). While these terms characterise groups of texts that share a certain property, these groupings cannot be considered genres themselves, rather, they are, in essence, groups of genres: "academic style" comprises the genres that are commonly found in academia, such as, for example, *M. A. thesis*, *technical report*, *review*, or *scholarly journal*. In contrast, "journalistic style" comprises all journalistic genres, such as *feature article*, and *news article*, probably even *weather report*, *letter to the editor*, and *obituary*. Also problematic are general text types such as "informational", "narrative", "argumentative" etc. While some researchers refer to categories such as these as "super-genres", they are also loose groupings of texts based on their respective purpose and their most salient discourse structure.

We would like to be able to represent both functional styles and generic text types of the level of abstraction sketched above in our reference corpus of web genres. As we plan to use stand-off annotation (see section 5.3), it is possible to prepare arbitrary groupings of genres in order to map, for example, a certain set of genres onto "academic style" and another set onto "official style". Similarly, arbitrary text properties or text types such as "informational", "instructional", or "narrative" can be either taken from existing genre assignments or annotated directly.

We also recognise a pressing need for detailed guidelines that help the annotators applying and, if none of the existing labels fit, extending the set of genre categories. For this purpose, the annotation guidelines should, for each category, include several example documents (or they should provide easy access to documents already annotated), and they should include precise definitions and instructions how to come up with new genre labels that have both adequate names and operate on an appropriate level of categorisation. We are currently discussing the advantages and disadvantages of annotation guidelines for this task (Rehm, 2008).

5.2. A Reference Collection of Documents

We plan to build the reference corpus in two stages. First, we will apply the category sets that we are currently working on (see section 5.1) to existing collections as a proof of concept. This can be thought of as a first step towards an objective evaluation and integrative comparability of individual approaches for automatic web genre identification. Second, we will use a web crawler to gather more recent as well as more diverse sets of documents. The annotation will be carried out by as diversified a group of web users as possible so that real users (in contrast to the researchers themselves exclusively) construct this crucial part of the resource; inter-coder reliability will be taken into account. Among the collections that we plan to process initially are:

- The Web Corpus for English (Santini, 2007), including (a) *editorials*, *short biographies*, *DIY mini-guides* and *feature articles* (20 web pages per category); (b) seven novel web genres annotated with objective sources (Santini, 2006); (c) the SPIRIT collection (Sanderson and Joho, 2004), which contains random and unclassified web pages.
- The Hierachical Webgenre Collection (Stubbe and Ringlstetter, 2007; Stubbe et al., 2007b), containing 32 genre classes, 40 HTML files per class, in English, collected in 2005/2006.
- The 20-Genre Collection (Vidulin et al., 2007).
- The Corpus of 400 blog posts (Tavosanis, 2007).
- The English and Russian web genre corpora (Sharoff, 2007a), including manually checked samples of 250 pages for each of the two languages, as well as predicted classes produced by SVM-based classifiers (65,177 pages for English, 29,650 for Russian).
- The German corpus by (Mehler et al., 2008) and (Mehler et al., 2007) including four web genres: 50 *conference websites* (2,779 pages, 435 annotated page segments), 68 *personal academic homepages* (1,569 pages, 292 segments), 52 *project websites* (1,591

1. **About Page** – A web page that presents personal or institutional information.
2. **Abstract** – Title and brief description or summary of the content.
3. **Agenda** (Schedule, Calendar) – List of upcoming or regular events, usually sorted by date and time.
4. **Announcement** – Announces an upcoming event.
5. **Application** – A web *application* (versus a web *document*).
6. **Bibliography** – List of books, journal articles or other publications.
7. **Biography** – Portrait of a person or organisation, usually written in prose.
8. **Chronicle** – A detailed and continuous register of events in order of time; a historical record (OED).
9. **Code Listings** – Source Code.
10. **Column / Editorial / Lead Article** – Personal opinion expressed by an author, or editor on a current topic; often appearing within a series.
11. **Comic**
12. **Contact Form** – Form for asking questions, sending comments and alike.
13. **Contract / Disclaimer / Terms and Conditons** – Exchange of promises or agreement (meant to be legally binding).
14. **Corporate Blog / Clog** – Blog run by a company or members of a company. Often on a specific topic, containig almost no personal stories.
15. **Curriculum Vitae / CV / Resume** – Usually written by the person in question; summary of personal career.
16. **Data / Statistics / Data Sheet** – Information presented mainly using numbers, and tables.
17. **Diary, Blog** – Personal narrative or time log of activities.
18. **Dictionary**
19. **Directory of Persons or Organisations** – List of the inhabitants of any locality, with their addresses and occupations; also a similar compilation dealing with the members of a particular profession, trade, or association (OED).
20. **Discussion Group / Newsgroup** – Discussion on a specific topic.
21. **Download** – Links to non-HTML files.
22. **Drama / Play**
23. **Encyclopedia** – Compendium that contains information on all branches of knowledge or a particular branch of knowledge or an article therein.
24. **Errata**
25. **Error Message / Empty Page / Under Construction Page**
26. **Essay** – Argumentative text on a specific topic expressing the personal opinion of the author.
27. **Exercises** (Problems)
28. **FAQ**
29. **Feature Story / News Reportage** – A longer article that takes an in-depth look at a subject (Wikipedia).
30. **Game** (Quiz, Puzzle)
31. **Glossary** – List of definitions.
32. **Guestbook**
33. **Homepage / Front Page / Entry Page** – The first page of a website.
34. **Horoscope**
35. **Index** – Web page with lots of links to the same website.
36. **Instruction** – Explains how to do something step by step.
37. **Interview**
38. **Invitation**
39. **Job Listing**
40. **Joke**
41. **Law / Regulation / Rule / Proclamation** – Adminstrative, regulatory texts.
42. **Letter / Mail / E-Mail** – Personal one-to-one communication.
43. **Letter to the Editor**
44. **Linkfarm** – Generated to attract traffic from web crawlers.
45. **Link Collection / Hotlist** – List of links to other web pages.
46. **List of Products**
47. **List of Projects**
48. **Login Page**
49. **Media** – Images, videos, music, sound.
50. **Meeting minutes**
51. **News Article**
52. **News Collection / Newsletter / Digest**
53. **Obituary**
54. **Official Report** – Formal statement of the results of an investigation or of any matter on which definite information is required (OED).
55. **Ordering Form / Booking Form**
56. **Pamphlet** – Small treatise on some subject or question of current interest, personal, social, political, ecclesiastical, or controversial, on which the writer desires to appeal to the public (OED).
57. **Petition** – Request to an authority.
58. **Promotional / Advertisement** – Presentation of institutions, products and services, pages of "institutionalized" individuals (movie stars, singers etc.).
59. **Poem / Poetry / Lyrics**
60. **Pornographic**
61. **Prose Fiction**
62. **Quotation**
63. **Reportage** – Longer story about travels, persons, events.
64. **Research Report**
65. **Review** (Testimonial) – Description and Evaluation.
66. **Script** (Manuscript)- The typescript of a cinema or television film (OED).
67. **Search Form**
68. **Sermon**
69. **Shop**
70. **Specification** – Describes some product or service in detail.
71. **Speech**
72. **Splash Page / Gateway / Welcome Page** – Introductory and (non automatic) redirection pages that take the visitor someplace else.
73. **Strategic Plans** – Actions to be taken in the future.
74. **Survey**
75. **Table of contents / Sitemap / Navigation** – A summary of the matters contained in a website.
76. **Thesis**
77. **Travel Guide**
78. **Tutorial** – Text, such as a school book, that explains something.

Table 4: An initial list of web genres compiled from previous approaches in a wiki-based discussion

pages, 612 segments), 180 *city websites* (based on 39,862 pages). Further, the English corpus built by (Mehler et al., 2007) including two web genres: the genre of 1,460 *conference websites* (76,011 pages), and *personal academic homepages* (16,652 pages). These corpora are built on the level of websites.

5.3. Corpus Management and Annotation Tools

The compilation and annotation of a reference corpus of web genres requires tools that operate on various levels of web documents. HyGraph is such a tool box and comes with a graphical user interface. It allows for the construction, storage, management, and retrieval of large corpora of web documents (Gleim et al., 2007) and has been designed to support researchers in the overall process of corpus compilation, annotation and analysis. Following, we describe how some of the typical steps of building corpora of web documents can be accomplished using HyGraph.

The first step concerns the extraction of web pages or entire websites. HyGraph includes a configurable crawler which downloads all reachable web pages based on a seed URL (see figure 2). The extent to which a crawler should follow hyperlinks to capture all documents of a website is a non-trivial task. The crawler can be configured to adhere to

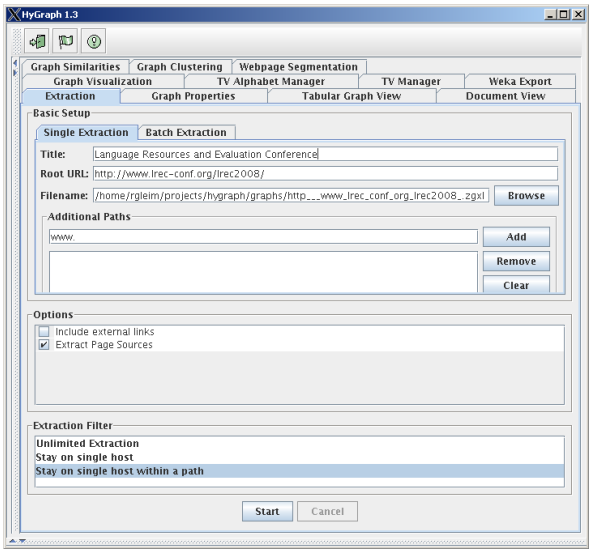

Figure 2: The HyGraph web crawler

predefined and extensible heuristics in order to determine the boundaries of a website at extraction time. The usual approach to building a corpus of a web genre (such as, in

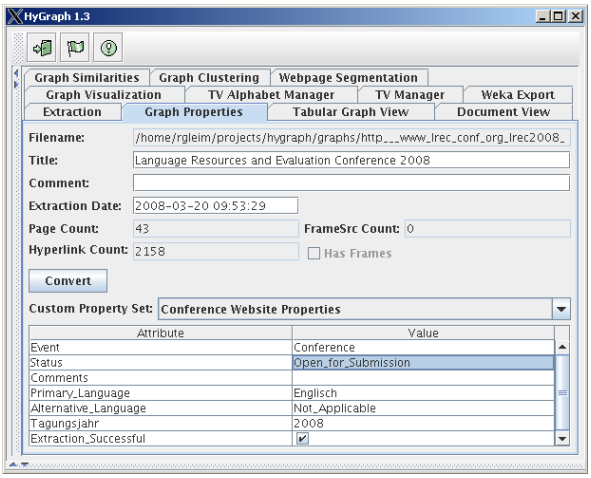

Figure 3: Annotating a web document

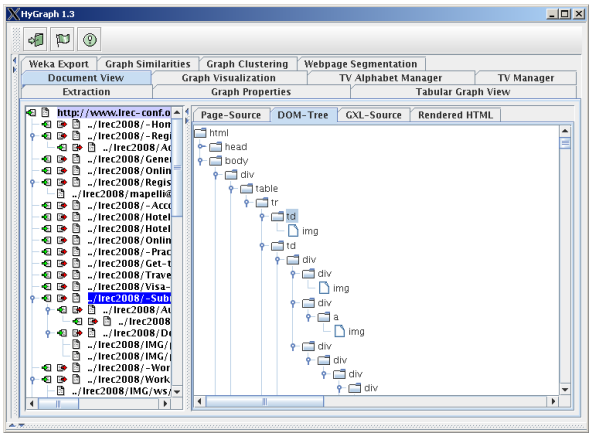

Figure 5: DOM view of an HTML document

our example, conference websites) would be to collect a list of URLs, possibly from a conference index, and then to use HyGraph's batch mode to extract all corresponding websites. The output of this process is a directory that contains all downloaded resources. Furthermore, the web document graph which is induced by the hyperlinks is extracted and stored in a dedicated XML file using GXL, the Graph eXchange Language (Holt et al., 2006). The graph representation distinguishes inter- and intra-page linking and also performs a basic typing of hyperlinks to capture the hierarchical structure of web documents and their components (Mehler and Gleim, 2006). The resulting GXL-file contains metadata and is also used to store any further annotation. The use of stand-off annotation leaves the original resources untouched and easily accessible for other tools or alternative approaches at grouping existing annotations.

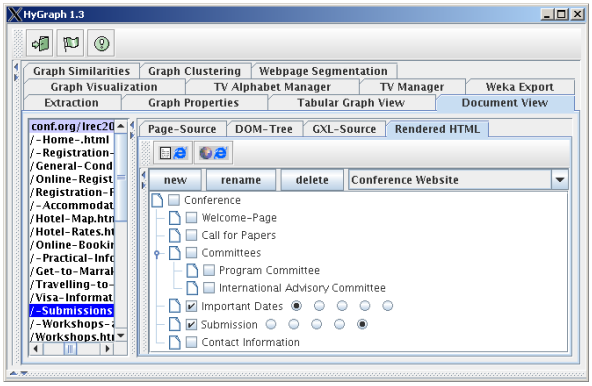

Figure 4: The dynamic categorisation module

The next step is the annotation of the extracted documents. Since the annotation is arguably *the* most crucial factor for the success of further analysis, the system puts special emphasis on it. Annotation can be done on various hierarchical levels. The most general one regards the entire document. Since different genre analyses ask for different, possibly structured tag sets, the latter can be freely configured. In the case of the web genre *conference website* we choose to annotate, amongst others, the type of event (e. g., *conference*, *workshop*, *symposium*), its state (e. g., *announced*, *open for submission* or *closed*), the primary language and the date of download (see figure 3). The next, more fine-grained level of annotation concerns the document level. In order to analyse corpora and to use resources as training data for machine learning algorithms it is crucial to annotate categorial information. The process of applying or extending an existing category system and then annotating pages or page segments with category labels is often interwoven and cannot be clearly separated. The need for new categories or using variant names, for example, often arises during this process. Therefore, HyGraph allows for the construction and annotation of dynamic category systems that evolve during the annotation process. Further, the degree to which a certain category label applies to a page can be specified by setting a confidence value ranging from one to five (see figure 4). Finally, a rendered view of resources is available in HyGraph itself and in external browsers (e. g., Firefox).

Figure 6: Website visualisation

Note that in many cases only segments of a page (vs. the whole page) manifest a specific category. Moreover, categorising web documents suffers from heterogeneous, i. e., polymorphic web pages that contain different patterns as instances of multiple categories (Mehler et al., 2007). Therefore, the possibility of explicitly demarcating subtrees of the DOM-tree of a page is an important part of annotating web documents. This functionality is currently under development; its inclusion in HyGraph will allow for annotating entire pages as well as DOM-based segments, see also (Rehm, 2005). HyGraph offers various means of exploring extracted web documents. It is possible to view the HTML source of web pages, their DOM-trees (figure 5) or their GXL-based representation. A graphical viewer for rapid website skimming is also included. Figure 6 shows an example visualisation of the LREC 2008 conference website. Finally, HyGraph is able to make managed web documents accessible to other tools and further analyses. While the re-

sources can be accessed via the file system, the GXL representation offers additional metadata and other, highly structured information that can be parsed by tools that work on GXL-based graph representations. HyGraph also supports the export of GXL files into the format of machine learning tools such as, for example, SVMLight or LibSVM.

6. Conclusions and Future Work

In this position paper we present a project that aims at constructing a reference corpus of web genres as a shared resource for researchers who work on automatic web genre identification approaches and the evaluation of these systems. Future work includes the realisation of this resource. We will start by applying a set of genre categories to existing corpora of web genres, whereupon we will collect a very large set of new documents that will be categorised based on detailed annotation guidelines using the HyGraph tool; legal issues will be taken into account, see (Grimmelmann, 2007; Lehmberg et al., 2008). While we will start by applying genre labels to complete HTML documents, we plan to apply similar category sets to page segments as well as to complete websites or hypertexts. We also plan to include functions for a monitor corpus so that we can observe and take into account how HTML documents and web genres change over time.

7. References

K. Adamzik. 1995. *Textsorten – Texttypologie. Eine kommentierte Bibliographie*. Nodus, Münster.

L. Björneborn. 2008. Genre Connectivity and Genre Drift in a Web of Genres. In A. Mehler, S. Sharoff, G. Rehm, and M. Santini, editors, *Genres on the Web*. In preparation.

E. S. Boese. 2005. Stereotyping the web: Genre classification of web documents. Master's thesis, Computer Science Department, Colorado State University.

A. Brandl. 2002. *Webangebote und ihre Klassifikation – Typische Merkmale aus Experten- und Rezipientenperspektive*. R. Fischer, München.

P. Braslavski. 2007. Combining Relevance and Genre-Related Rankings: An Exploratory Study. In G. Rehm and M. Santini, editors, *Proc. of the Int. Workshop Towards Genre-Enabled Search Engines*, pages 1–4.

M. Dimter. 1981. *Textklassenkonzepte heutiger Alltagssprache – Kommunikationssituation, Textfunktion und Textinhalt als Kategorien alltagssprachlicher Textklassifikation*, volume 32 of *Reihe Germanistische Linguistik*. Niemeyer, Tübingen.

L. Dong, C. Watters, J. Duffy, and M. Shepherd. 2008. An Examination of Genre Attributes for Web Page Classification. In *Proc. of the 41st Hawaii Int. Conf. on Systems Sciences (HICSS-41)*.

A. Ferrari and E. Manzotti. 2002. Linguistica del testo. In C. Lavinio, editor, *La linguistica italiana alle soglie del 2000 (1987–1997 e oltre)*, pages 413–453. Bulzoni, Roma.

A. Finn and N. Kushmerick. 2006. Learning to Classify Documents According to Genre. *Journal of the Am. Soc. for Inf. Science and Tech.*, 57(11):1506–1518.

R. Gleim, A. Mehler, and H.-J. Eikmeyer. 2007. Representing and Maintaining Large Corpora. In *Proc. of Corpus Linguistics 2007*, Birmingham, UK.

J. Grimmelmann. 2007. The Structure of Search Engine Law. *Iowa Law Review*, 93(1):1–63.

S. Gupta, H. Becker, G. Kaiser, and S. Stolfo. 2006. Verifying genre-based clustering approach to content extraction. In *Proc. of the 15th Int. Conf. on World Wide Web*, pages 875–876. ACM Press.

R. C. Holt, A. Schürr, S. Elliott Sim, , and A. Winter. 2006. GXL: A Graph-Based Standard Exchange Format for Reengineering. *Science of Computer Programming*, 60(2):149–170.

M. Y. Ivory and M. A. Hearst. 2002. Statistical Profiles of Highly-Rated Web Sites. In *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*, pages 367–374, Minneapolis. ACM Press.

J. Karlgren and D. Cutting. 1994. Recognizing Text Genres with Simple Metrics Using Discriminant Analysis. In *COLING 94 – The 15th Int. Conf. on Computational Linguistics*, volume 2, pages 1071–1075, Kyoto, August. ACL.

J. Karlgren, I. Bretan, J. Dewe, A. Hallberg, and N. Wolkert. 1998. Iterative Information Retrieval Using Fast Clustering and Usage-Specific Genres. In *Proc. of the 8th DELOS Workshop on User Interfaces in Digital Libraries*, pages 85–92.

A. Kennedy and M. Shepherd. 2005. Automatic Identification of Home Pages on the Web. In *Proc. of the 38th Hawaii Int. Conf. on Systems Sciences (HICSS-38)*.

B. Kessler, G. Nunberg, and H. Schütze. 1997. Automatic Detection of Text Genre. In *Proc. of the 35th Annual Meeting of the ACL*, pages 32–38.

Y. Kim and S. Ross. 2007a. Searching for Ground Truth: A Stepping Stone in Automated Genre Classification. In Thanos et al., editor, *Proc. of the DELOS Conf. on Digital Libraries*, pages 248–261.

Y. Kim and S. Ross. 2007b. Variations of Word Frequencies in Genre Classification Tasks. In *Proc. of the DELOS Conf. on Digital Libraries*, Tirrenia, Italy.

Y.-B. Lee and S. Hyon Myaeng. 2004. Automatic Identification of Text Genres and Their Roles in Subject-Based Categorization. In *Proc. of the 37th Hawaii Int. Conf. on Systems Sciences (HICSS-37)*.

T. Lehmberg, G. Rehm, A. Witt, and F. Zimmermann. 2008. Preserving Linguistic Resources: Licensing – Privacy Issues – Mashups. *Library Trends*. In print.

R. Levering, M. Cutler, and L. Yu. 2008. Using Visual Features for Fine-Grained Genre Classification of Web Pages. In *Proc. of the 41st Hawaii Int. Conf. on Systems Sciences (HICSS-41)*.

C. Su Lim, K. Joo Lee, and G. Chang Kim. 2005. Multiple Sets of Features for Automatic Genre Classification of Web Documents. *Information Processing and Management*, 41(5):1263–1276.

L. Littig and C. Lindemann. 2008. Classification of Web Sites at Super-Genre Level. In A. Mehler, S. Sharoff, G. Rehm, and M. Santini, editors, *Genres on the Web*. In preparation.

A. Mehler and R. Gleim. 2006. The Net for the Graphs – Towards Webgenre Representation for Corpus Linguistic Studies. In M. Baroni and S. Bernardini, editors, *WaCky! Working Papers on the Web as Corpus*, pages 191–224. Gedit, Bologna.

A. Mehler, R. Gleim, and A. Wegner. 2007. Structural Uncertainty of Hypertext Types. In G. Rehm and M. Santini, editors, *Proc. of the Int. Workshop Towards Genre-Enabled Search Engines*, pages 13–20.

A. Mehler, U. Waltinger, R. Gleim, and A. Wegner. 2008. A Model of Semi-Supervised Hypertext Zoning. In A. Mehler, K.-U. Kühnberger, H. Lobin, H. Lüngen, A. Storrer, and A. Witt, editors, *Modeling, Learning and Processing of Text Technological Data Structures*. In preparation.

A. Mehler. 2008. Structural Similarities of Complex Networks: A Computational Model by Example of Wiki Graph. *Applied Artificial Intelligence*. In print.

S. Meyer zu Eissen and B. Stein. 2004. Genre Classification of Web Pages. In *Proc. of the 27th German Conf. on AI (KI-2004)*, Ulm, September.

G. Rehm and M. Santini, editors. 2007. *Proc. of the Int. Workshop Towards Genre-Enabled Search Engines: The Impact of NLP*, Borovets, Bulgaria.

G. Rehm. 2002. Towards Automatic Web Genre Identification – A Corpus-Based Approach in the Domain of Academia by Example of the Academic's Personal Homepage. In *Proc. of the 35th Hawaii Int. Conf. on System Sc. (HICSS-35)*.

G. Rehm. 2005. Language-Independent Text Parsing of Arbitrary HTML-Documents – Towards A Foundation For Web Genre Identification. *LDV Forum*, 20(2):53–74.

G. Rehm. 2007. *Hypertextsorten: Definition – Struktur – Klassifikation*. Books on Demand, Norderstedt. (PhD thesis, Comp. Ling., Giessen University, 2005).

G. Rehm. 2008. A Comparative Analysis of Genre Category Sets as a Prerequisite for a Reference Corpus of Web Genres. In A. Mehler, S. Sharoff, G. Rehm, and M. Santini, editors, *Genres on the Web*. In preparation.

M. Rosso. 2005. *Using Genre to Improve Web Search*. Ph. D. thesis, School of Inf. and Lib. Sc., Univ. of North Carolina at Chapel Hill.

M. Rosso. 2008. User-Based Identification of Web Genres. *JASIST*, 59(5):1–20.

M. Sanderson and H. Joho. 2004. Forming test collections with no system pooling. In K. Järvelin, J. Allan, P. Bruza, and M. Sanderson, editors, *Proc. of the 27th Int. ACM SIGIR Conf. on Research and Dev. in IR*, pages 33–40.

M. Santini. 2006. Common Criteria for Genre Classification: Annotation and Granularity. In *Workshop on Text-based IR (TIR 06)*, Riva del Garda, Italy.

M. Santini. 2007. *Automatic Identification of Genre in Web Pages*. Ph.D. thesis, University of Brighton.

M. Santini. 2008. Zero, Single, or Multi? Genre of Web Pages through the Users' Perspective. *Information Processing and Management*, 44(2):702–737.

S. Sharoff. 2007a. Classifying Web Corpora into Domain and Genre Using Automatic Feature Identification. In *Proc. of Web as Corpus Workshop*, Louvain-la-Neuve, September.

S. Sharoff. 2007b. In the Garden and in the Jungle: Comparing Genres in the BNC and Internet. In M. Santini and S. Sharoff, editors, *Proc. of the Colloquium Towards a Reference Corpus of Web Genres*, Birmingham, UK, July.

A. Stubbe and C. Ringlstetter. 2007. Recognizing Genres. In M. Santini and S. Sharoff, editors, *Proc. of the Colloquium Towards a Reference Corpus of Web Genres*, Birmingham, UK, July.

A. Stubbe, C. Ringlstetter, and R. Goebel. 2007a. Elements of a Learning Interface for Genre Qualified Search. In G. Rehm and M. Santini, editors, *Proc. of the Int. Workshop Towards Genre-Enabled Search Engines*, pages 21–28.

A. Stubbe, C. Ringlstetter, and K. U. Schulz. 2007b. Genre to Classify Noise – Noise to Classify Genre. In *Proc. of the IJCAI-2007 Workshop on Analytics for Noisy Unstructured Text Data*, Hyderabad, India.

S. Symonenko. 2007. Recognizing Genre-Like Regularities in Website Content Structure. In G. Rehm and M. Santini, editors, *Proc. of the Int. Workshop Towards Genre-Enabled Search Engines*, pages 29–36.

M. Tavosanis. 2007. Juvenile Netspeak and Subgenre Classification Issues in Italian Blogs. In G. Rehm and M. Santini, editors, *Proc. of the Int. Workshop Towards Genre-Enabled Search Engines*, pages 37–43.

V. Vidulin, M. Luštrek, and M. Gams. 2007. Using Genres to Improve Search Engines. In G. Rehm and M. Santini, editors, *Proc. of the Int. Workshop Towards Genre-Enabled Search Engines*, pages 45–51.

J. Xu, Y. Cao, H. Li, N. Craswell, and Y. Huang. 2007. Searching Documents Based on Relevance and Type. In G. Amati, C. Carpineto, and G. Romano, editors, *Proc. of the 29th European Conf. on IR Research (ECIR 2007)*, pages 629–636.

P.C.K. Yeung, S. Büttcher, C.L.A. Clarke, and M. Kolla. 2007. A Bayesian Approach for Learning Document Type Relevance. In G. Amati, C. Carpineto, and G. Romano, editors, *Proc. of the 29th European Conf. on IR Research (ECIR 2007)*, pages 753–765.