Marina Santini, Serge Sharoff

Web Genre Benchmark Under Construction

The project discussed in this article focuses on the creation of web genre benchmarks (a.k.a. web genre reference corpora or web genre test collections), i.e. newly conceived test collections against which it will be possible to judge the performance of future genre-enabled web applications. The creation of web genre benchmarks is of key importance for the next generation of web applications because, at present, it is impossible to evaluate existing and in-progress genre-enabled prototypes. We suggest focusing on the following key points: 1) propose a characterisation of genre suitable for digital environments and empirical approaches shared by a number of genre experts working in automatic genre identification; 2) define the criteria for the construction of web genre benchmarks and draw up annotation guidelines; 3) create several web genre benchmarks in several languages; 4) validate the methodology and evaluate the results.

1 The Concept of Genre

The concept of genre is hard to agree upon. Many interpretations have been proposed since Aristotle's Poetics (who was mostly discussing literary genres, such as epic, lyric, and drama) without any definite conclusions about the inventory or even principles for classifying documents into genres. Some studies put the number of genres to 4,500 (Adamzik, 1995). Recently definitions of genre have been adapted to the new digital environments, e.g., (Yates and Orlikowski, 1992; Erickson, 1999; Toms and Campbell, 1999; Beghtol, 2001; Heyd, 2008; Bateman, 2008).

Researchers are working with genres of electronic documents, such as FAQs, e-shops, home pages, or conference websites in order to better satisfy users' needs in a number of different application areas, such as Information Retrieval e.g., (Stamatatos et al., 2000; Meyer zu Eissen and Stein, 2004), digital libraries e.g., (Rauber and Müller-Kögler, 2001; Kim and Ross, 2009), and information extraction e.g., (Maynard et al., 2001; Gupta et al., 2006).

The lack of an agreed definition of what genre is causes the problem of the loose boundaries between the term 'genre' with other neighbouring terms, such as 'register', 'domain', 'topic', 'style'. The inventory of genres can be based on linguistic theories or 'folksonomies', i.e. labels used by users (e.g. users are confident with a term like *novel*), whereas linguistic researchers may prefer functional terms, like *recreation* to indicate a bit wider range of genres aimed at recreational reading. Undoubtedly, the situation on the web is even more difficult than in the offline world, because the Web is new, genres are fluid, web documents are very often characterised by a high level of hybridism, by

the fragmentation of textuality across several documents, by the impact of technical features such as hyperlinking and posting facilities.

Nevertheless, as stressed by Karlgren (2005) the term 'genre' is established and generally understood, at least intuitively, and it is currently used in many real-world environments. For instance, online bookshops, like Amazon, organise their catalogues by genre.[1]

2 Automatic Genre Identification

Traditionally, scholars and researchers studying the genre of documents annotate these documents themselves, i.e. manually. The main drawback with manual annotation is that it is extremely tedious and time-consuming. Consequently, the number of documents manually annotated by genre is often too small to have a full picture of certain phenomena or to carry out any quantitative approach. Additionally, now with the web and with the wealth of freely available documents, the 'manual annotation pace' is certainly a huge limitation for genre research. The second drawback is that since manual annotation is a mentally demanding activity, tiredness or distraction causes errors and idiosyncrasies. Ideally, as machines do not get tired, they should provide genre analysts with larger quantity of consistently genre-annotated documents. In brief, annotating documents by genre is not always an easy task: it takes time, it is not always intuitive and it is prone to errors, because human annotators get easily tired or confused. For this reason, automatic genre identification (AGI) would be a great advantage.

Attempts at automatic genre identification start from (Karlgren and Cutting, 1994; Kessler et al., 1997) both applied to texts from the Brown Corpus. The first prototype of a genre-enabled application for the web was created in 1998 (Karlgren et al., 1998) (see DropJaw below). More recently, a genre add-on that can be installed on to a general-purpose search engine (namely Mozilla Firefox) has been completed at Bauhaus University Weimar, Germany (Stein et al., 2009) (see WEGA below). In both cases, these applications could not and cannot be fully evaluated because of the absence of web genre benchmarks enabling the objective assessment of their effectiveness.

Web genre benchmarks are still missing because their design and construction is difficult. So far, many national and 'ad-hoc' corpora have been built to represent the language, but very few large corpora indicate the genres of the documents they include, and when they do, classifications are not consistent. For example, there are several competing genre-related classifications available in the British National Corpus (BNC), such as the publication medium (book, periodical, etc), audience level, as well as a set of 70 labels called *genres*, such as 'academic texts in social sciences' (Lee, 2001). 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

[1] `http://www.amazon.co.uk/Books-Categories/b?ie=UTF8&node=1025612`

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.

Not happy with the genres included in these kinds of corpora, many researchers have created their own genre collections with their own inventories of genre categories. Some researches have created a hierarchy where super-genres are broken down to different medium-level genre classes, e.g., (Stubbe and Ringlstetter, 2007). Others have used more general categories such as the functional styles of the Russian linguistic tradition derived from the Prague Linguistic Circle, e.g. *everyday* or *journalistic* (Braslavski, 2009), or functional classes derived from the corpus-linguistic tradition, e.g. *instructional* or *recreation* (Sharoff, 2009).

While many current genre collections have the individual web pages as unit of analysis, another line of genre research focuses on genre classes at web site level. For instance, Symonenko (2007) identifies genre-like regularities in the content structure in commercial and educational websites; Rehm (2002) analyses the genre of academic personal homepages, while Mehler et al. (2007) focuses on city websites, conference websites, and personal academic homepages.

In this thriving of genre classes and genre corpora assembled with interest-specific criteria, a practice has established very recently, namely the testing of classification models over several existing web genre collections. This *cross-testing* technique has been adopted by Santini (2009), Kim and Ross (2009), Kanaris and Efstathios (2007) and others. This practice represents a step forward, but only partially addresses the issues underlying the need for a more objective assessment of genre classes, because existing genre collections have been built without the ambition of being genre benchmarks, consequently they do not have the requirements for being a "reference" or a "standard".

Genre-Enabled Prototypes The existence of a number of genre-enabled prototypes show the potential of genre in real-world applications. Notably, four prototypes have been described and documented, namely: DropJaw, Hyppia, X-Site and WEGA.

DropJaw: Karlgren and co-workers (Karlgren et al., 1998) built a fully functional prototype system, DropJaw, to experiment with iterative search. They use World Wide Web, and DropJaw bases its searches for web documents by the user entering terms, as in a traditional system. However, rather than producing ranked lists of output based on term occurrence, DropJaw displays the distribution of the resulting set over two dimensions: dynamically generated topical clusters and user-defined, document-base oriented genre. The two-dimensional document space is displayed on a work board or matrix for further user processing.

Hyppia: (Dimitrova and Kushmerick, 2003) describe how shallow text classification techniques can be used to sort the documents returned by web search engines according to genre dimensions, such as: the degree of expertise assumed by the document; the amount of detail presented; whether the document reports mainly facts or opinions. (Dimitrova and Kushmerick, 2003) are connected to the Hyppia project. The Hyppia demo allows news articles to be filtered and searched based on genre information. The

genre classes in this demo are considered to be whether a document is subjective or objective (Finn et al., 2002; Finn and Kushmerick, 2006).

X-SITE: X-Site is a search system designed and implemented to test the practical value of making use of task-genre relationships in a real life work environment (Freund, 2008). X-Site was implemented as an extension to MultiText, a pre-existing indexing and retrieval engine (Freund, 2008). X-Site makes use of three contextual components in addition to the basic search engine functionality, namely 1) a genre classifier, which uses ML methods; 2) a task profile, which composed of a work task and an information task; and 3) a task-genre association matrix, which specifies the relationships between the task taxonomies and the genre taxonomies.

WEGA: While X-Site has been devised for professionals (namely software engineers) who can exploit the concept of genre to rapidly find information that is task-appropriate, situationally-relevant and mission-critical for their job, WEGA (an acronym that stands for WEb Genre Analysis), (Stein et al., 2009) has been designed for the web and for common web users. WEGA is an add-on that superimposes genre labels a few seconds after the result list is returned by a general-purpose search engine, namely Mozilla Firefox.

These prototypes show that genre-enabled systems are feasible[2] and that genre classes can help improve productivity in the workplace (in the case of X-Site) and offer additional hints about the nature of the web pages listed in the search results (in the case of WEGA). The next step is then to provide an evaluation resources to test these applications. The design and construction of genre-annotated resources is also very timely since genre-enabled applications are a hot topic in current research, e.g., (Mehler et al., 2009).

In this article we outline a project for the creation of web genre benchmarks, against which it will be possible to judge the performance of genre-enabled web applications.

3 Research Goals

The major research efforts for the creation of web genre benchmarks are devoted to:

1. propose a characterisation of genre suitable for digital environments and empirical approaches;

2. define the criteria for the construction of genre benchmarks and draw up annotation guidelines;

3. create genre benchmarks in several languages;

4. validate the methodology and evaluate the results.

[2] Additionally, a number of patent application has been submitted in the United States by XEROX Corporation on the basis of work from (Kessler et al., 1997).

The aim is to enable the comparison of different empirical approaches, to objectively evaluate the performance of different computational methodologies, and, last but not least, to assess the impact of the number of genres, the number of documents, the number of annotators and the criteria of annotation may have on genre findings and on the performance of genre-enabled applications.

4 Genre classes

The construction of genre benchmarks necessarily involves the task of assigning motivated labels to documents. Although the term genre can be intuitively understood, huge problems arise when it comes to the identification of which document classes can be considered "genres". While the Sidney School, centred upon the Systemic Functional Linguistics, e.g., Martin and Rose (2008), focuses more on the role of genre in the linguistic communication system, the North American School e.g., (Swales, 1990; Yates and Orlikowski, 1992) focuses on the genres used in specific communities, e.g., Swales accounts for research and academic genres. The German text linguistics school established tradition of cataloguing genre classes (called Textsorten), e.g., Görlach (2004) counted about 2000 genres, or text types, based on Shorter Oxford English Dictionary.

However, these nomenclatures or taxonomies seem to be disconnected from the current trends in automatic genre identification, which is currently handling a proliferation of classes that are not, properly speaking, genres. Some of them have been created in ad-hoc fashion (e.g. *tables* or *lists, person, resources, children, subjective opinion, content delivery*, etc.) because they are assumed to be useful classes when searching the web. An interesting discussion on this point can be found in (Karlgren, 2009), where the author suggests that it not enough to discover new surface features to postulate new genres. On the contrary, it is the study of information needs that allow us to detect them, since genres are behavioural categories.

Therefore, one challenge in the construction of genre benchmarks is to convey the variety of genre classes that have been used so far in automatic genre classification experiments without cutting out others that can be potentially useful for the information needs of web users (Issue 1). The inclusion of different variety of genre results in genre-diversified benchmarks, which will permit to test the exportability of genre classification systems built on a set of genre classes to a different set (Issue 2). However, it is hard to devise benchmarks that are easily updated with the new genres brought about the advances of web technology (Issue 3). Another challenge is represented by the size of the benchmarks: although designed to be large, benchmark corpora are necessarily limited in size. What is the minimum corpus size (or critical mass) required to test the scalability of genre-enabled applications (Issue 4)? Albeit genre colonisation (Beghtol, 2001) is quite extensive on the web, genre classes are social artefacts linked to specific cultures (e.g. it seems that the *obituary* genre is not indigenous to Asian countries), so one must decide about the cross-cultural span of the genre benchmarks (Issue 5). We would like to address these problems as follows:

Issue 1 Diversified genre palettes will be included in the benchmarks thus allowing a large diversification of genre classes.

Issue 2 This variety of genres is useful to test the exportability of genre classification systems.

Issue 3 We plan to monitor genre evolution through a monitor corpus. As the construction of genre monitor corpora is not a trivial issue and has the creation of genre benchmarks as prerequisite, we postpone its creation to future research and projects.

Issue 4 Web genre corpora of different sizes will be devised to investigate problems related to scalability.

Issue 5 Development of resources for several different languages will allow us to investigate the cultural distance (if any) between the cultures of different countries.

5 The Roadmap

5.1 Short-Term Plan: Mapping existing web genre collections into macro-genres

In the short term, the plan is to capitalise on existing genre-annotated resources. In this "small-scale" work plan we would like to re-utilise the web genre collections listed in Table 1[3]. They are all made of manually annotated English web documents, mostly in HTML.[4] Other discussions on genre-related issues can be found in `http://purl.org/net/webgenres`., by providing a stand-off annotation following the genre palette proposed in (Sharoff, 2009). This palette is compact and coarse-grained and we hypothesise that conflating finer-grained genre classes into coarser-grained functional genres may be more straightforward than mapping in the other direction.

We will create a shared format for storing these diverse collections and draw up guidelines for mapping.

In the end of this phase, all the documents of the six collections will be provided by their original genre annotation plus our stand-off annotation into seven classes. This means that we will have about 10.000 webpages annotated consistently. In this way, computational experiments can be carried out with seven genre classes and 10.000 web documents. We conjecture that significant insights will be yielded by the experiments performed on such a corpus.

In the second phase of the short-term plan, we would like to utilise the fine-grained palette presented in (Rehm et al., 2008). In this phase, our mapping scheme will have to accommodate three problematic cases:

[3] The collections listed in Table 1 are accessible from `http://www.webgenrewiki.org/index.php5/Genre_Collection_Repository`

[4] In addition to annotated pages, the SANTINIS corpus also contains pages without annotation considered as "noise" for the purposes of machine learning. KRYS-I collection contains PDF pages.

Table 1: Existing web genre collections used in the first phase of the short-term plan

Source	# pages	Genres
KI-04 (Meyer zu Eissen and Stein, 2004)	1205	8
SANTINIS (Santini, 2009)	2480	11
I-EN (Sharoff, 2009)	250	7
MGC (Vidulin et al., this issue)	1239	20
HGC (Stubbe and Ringlstetter, 2007)	1280	32
KRYS I (Berninger et al., 2008)	5305	70

1. a many-to-one mapping, when the target collection does not make finer-grained distinctions made in a source collection;

2. a one-to-many mapping, when the more general class of a source collection can be mapped into more than one genre class in our target palette;

3. a many-to-many mapping, when the two classification schemes are incompatible. For cases of many-to-many mappings between classes we will define additional features needed to achieve unambiguous mapping between individual documents.

Our hypothesis is that it should be possible to design a mapping on the level of classes in each collection, use automatic classification methods for approximate reclassification of more general classes and review their results manually.

5.2 Long-Term Plan

Phase I: Discussion, Decisions and Guidelines Building up on the experience accumulated during the short-term plan activities, we will start the long-term plan by building upon the 10,000 web document corpus. This stage will provide a flexible definition of web genre for computational purposes and a comprehensive annotation manual to reduce the level of ambiguity.

Phase II: Genre Benchmark Construction In this phase, the collection, annotation and storage of the web documents following the criteria defined in Phase I will take place. We anticipate that a number of genre corpora will be built during this phase. While the short term plan focused on English, in this phase we plan the construction of three corpora of web documents in several languages. Provisionally, we call these three corpora: "gold" corpus, "main" corpus and "comprehensive" corpus.

One "gold" corpus for each language will be annotated by several annotators to assist in studies measuring the level of disagreement between annotators, as well as cases of genre hybridism. With this smaller corpus we will also investigate the effect of using

radically different genre palettes, i.e. documents will be annotated with codes taken from incompatible sets of genres.

Documents in a "main" corpus for each language will be annotated, each following the main annotation schemes resulting from Phase I.

We will also prepare a "comprehensive" corpus (on the order of hundreds of thousands documents), which will be annotated automatically. We will train statistical classification models on the basis of the "main" corpus, leveraging on semi-supervised machine learning techniques (e.g. boostrapping and active learning) and apply them to the bigger corpus. With the "comprehensive" corpus, we would like to address two research issues:

1. genre hybridism, i.e. several separate genres in a single page, e.g., a newspaper article and a forum discussing it;

2. ambiguity in interpretation, e.g., ambiguity in the genre palette itself, see the description of wikipedia pages in (Rehm et al., 2008).

Importantly, all the corpora will follow a multi-labelling annotation scheme, where web pages are not necessarily (and artificially) restricted to the membership to a single genre. Techniques will be developed to establish sensible labelling thresholds. With the approach proposed above, web pages will be endowed by zero, one or more genre labels, as needed. This will allow future investigators to shed some light on whether the 'nature' of genre and the annotation method affect the performance of genre-enabled applications. Even if the quality of automatic classification in the "comprehensive" corpus could be far from being perfect, a really big genre-annotated corpus should help researchers in estimating the performance of their models on large-scale resources, one of the main holes in current automatic genre research.

Phase III: Evaluation of Machine Learning Techniques In this phase, the criteria and the experience built up in the previous phases will be used to develop reliable automatic genre classification models. During this phase, new evaluation methods and measure will be proposed to investigate the correlation among different genre granularity and classification schemes. Previous experiments have already shown that computable relations exist between rhetorical genres (like narration or argumentation) and social genres (such as blogs and editorials). For instance, see the two-layer approach proposed by Santini (2009), where these relations have been investigated only on small and heterogeneous genre corpora, which did not allow a robust evaluation of the results. The construction of principled benchmarks will allow us to delve deeply into evaluation techniques and eventually propose new evaluation measures, which more suitably account for classifiers' performance with difficult classes like genres. It is worth emphasising that multi-labelling genre evaluation is a challenging and very little explored field (an exception is Vidulin et al. in this Issue), and the contribution of this project in this respect will certainly be remarkable.

6 Corpus design issues

Since the web is a huge reservoir of texts that can be easily mined, we propose building genre benchmarks with freely downloadable web documents. This decision still leaves us with a range of open questions.

Document type The web is not limited to HTML pages, many documents are available in PDF, Word or PowerPoint files. Given that the set of affordances of non-HTML documents is considerably different from HTML pages, also methods and approaches to feature extraction can be specific for each document type, in the first version of our resources we propose to use only HTML and PDF pages.

Document selection The second open question concerns the criteria for selecting documents. Some researchers attempted to use equal amount of texts per genre, while others mined random samples of webpages for a given language or used existing text collections. This project is aimed at producing a set of diversified genre classes, thus resulting in multiple corpora corresponding to multiple benchmarks. In the end, the exact inventory of genres cannot be fixed and the corpus cannot be balanced by this criterion a priori. At the same time, a set of annotated texts from the total set of texts can be selected according to wishes of individual researchers, e.g. the subset chosen by a researcher can contain 200 news items vs. 100 editorials. The second argument in favour of using a random sample from the web for initial annotation is related to the purpose of our benchmarks, which have to reflect the composition of the web to be useful in application domains.

Copyright Another crucial question concerns the copyright. According to existing copyright law researchers are free to distribute URL links with their descriptions, from which it is possible to recreate a corpus in any necessary format (Sharoff, 2006b). The major problem with this method is that the Web changes, some pages get deleted, others updated. An experiment in measuring the decay rate of URLs estimated the half-life of an Internet corpus as about seven years, i.e. about half of the pages get changed or deleted in about seven years (Sharoff, 2006a). Storage and redistribution of complete webpages is not traditionally allowed under the copyright law. Some Internet corpus projects managed to overcome this constraint by putting sentences in their corpora in random order, for instance, some portions of the Hunglish corpus have been shuffled (Varga et al., 2007). This makes it possible to redistribute the content of webpages with appropriate annotations, but prevents investigation of any context larger than a sentence or doing discourse analysis. The most suitable solution for development of our reference webgenre corpus is to follow the practice of distribution of other webcorpora, such as deWac (Baroni and Kilgarriff, 2006) or ukWac (Ferraresi et al., 2008), which give the provision for copyright holders of individual webpages to opt out from keeping the their pages in the collection. In addition, it is possible to select webpages explicitly marked with permissive licences, such as the GNU Free Documentation Licence or a

family of Creative Commons Licences, even though this choice can bias the selection of texts.

7 Significance of the Research and Conclusion

This project will provide the community of genre scholars and practitioners with a number of theoretical contributions, and several valuable resources.

From a theoretical point of view, this project will enrich genre studies and genre research with a characterisation of the concept of genre tailored for digital environments. It will also produce a set of re-usable criteria for the construction of web genre benchmarks and annotation guidelines. Last but not least, it will provide long-lasting resources, namely a number of web genre benchmarks in several languages, which can be updated, monitored and enlarged in future.

References

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

Baroni, M. and Kilgarriff, A. (2006). Large linguistically-processed Web corpora for multiple languages. In *Companion Volume to Proc. of the European Association of Computational Linguistics*, pages 87–90, Trento.

Bateman, J. (2008). *Multimodality and genre: A foundation for the systematic analysis of multimodal documents*. Palgrave Macmillan.

Beghtol, C. (2001). The concept of genre and its characteristic. *Bulletin of ASIST*, 27(2):17–19.

Berninger, V., Kim, Y., and Ross, S. (2008). Building a document genre corpus: a profile of the KRYS I corpus. In *Proceedings of the Corpus Profiling Workshop*, London.

Braslavski, P. (2009). Marrying relevance and genre rankings: an exploratory study. In Mehler et al. (2009).

Dimitrova, M. and Kushmerick, N. (2003). Dimensions of web genre. In *World Wide Web Conference WWW2003, Budapest, Hungary*.

Erickson, T. (1999). Rhyme and punishment: the creation and enforcement of conventionsin an on-line participatory limerick genre. In *Proc. 32nd Annual Hawaii International Conference on System Sciences*.

Ferraresi, A., Zanchetta, E., Bernardini, S., and Baroni, M. (2008). Introducing and evaluating ukWaC, a very large web-derived corpus of English. In *The 4th Web as Corpus Workshop: Can we beat Google? (at LREC 2008)*, Marrakech.

Finn, A. and Kushmerick, N. (2006). Learning to classify documents according to genre. *Journal of the American Society for Information Science and Technology*, 57(11).

Finn, A., Kushmerick, N., and Smyth, B. (2002). Genre classification and domain transfer for information filtering. In *Proc. European Colloquium on Information Retrieval Research*, Glasgow.

Freund, L. (2008). *Exploiting task-document relationships to support information retrieval in the workplace*. PhD thesis, University of Toronto.

Görlach, M. (2004). *Text types and the history of English*. Walter de Gruyter.

Gupta, S., Becker, H., Kaiser, G., and Stolfo, S. (2006). Verifying genre-based clustering approach to content extraction. In *Proceedings of the 15th international conference on World Wide Web*, pages 875–876. ACM.

Heyd, T. (2008). *Email hoaxes: form, function, genre ecology*. Benjamins.

Kanaris, I. and Efstathios, S. (2007). Webpage genre identification using variable-length character n-grams. In *Proceedings of ICTAI*.

Karlgren, J. (2005). The whys and wherefores for studying textual genre computationally. In *Proc. AAAI Fall Symposium on Style and Meaning in Language, Art and Music*, Arlington, USA.

Karlgren, J. (2009). Conventions and mutual expectations — understanding sources for web genres. In Mehler et al. (2009).

Karlgren, J., Bretan, I., Dewe, J., Hallberg, A., and Wolkert, N. (1998). Iterative information retrieval using fast clustering and usage-specific genres. In *Eight DELOS workshop on User Interfaces in Digital Libraries*, pages 85–92.

Karlgren, J. and Cutting, D. (1994). Recognizing text genres with simple metrics using discriminant analysis. In *Proc. of the 15th. International Conference on Computational Linguistics (*COLING 94*)*, pages 1071 – 1075, Kyoto, Japan.

Kessler, B., Nunberg, G., and Schütze, H. (1997). Automatic detection of text genre. In *Proceedings of the 35^{th} ACL/8^{th} EACL*, pages 32–38.

Kim, Y. and Ross, S. (2009). Formulating representative features with respect to genre classification. In Mehler et al. (2009).

Lee, D. (2001). Genres, registers, text types, domains, and styles: clarifying the concepts and navigating a path through the BNC jungle. *Language Learning and Technology*, 5(3):37–72.

Martin, J. and Rose, D. (2008). *Genre Relations: mapping culture*. Equinox Pub.

Maynard, D., Tablan, V., Ursu, C., Cunningham, H., and Wilks, Y. (2001). Named entity recognition from diverse text types. In *Proc. Recent Advances in Natural Language Processing*, pages 257–274.

Mehler, A., Gleim, R., and Wegner, A. (2007). Structural uncertainty of hypertext types. an empirical study. In *Proc. Towards Genre-Enabled Search Engines: The Impact of NLP. RANLP-07*.

Mehler, A., Sharoff, S., Rehm, G., and Santini, M., editors (2009). *Genres on the Web: Computational Models and Empirical Studies*. Springer, Berlin/New York.

Meyer zu Eissen, S. and Stein, B. (2004). Genre classification of web pages. In *Proceedings of the 27th German Conference on Artificial Intelligence*, Ulm, Germany.

Rauber, A. and Müller-Kögler, A. (2001). Integrating automatic genre analysis into digital libraries. In *JCDL '01: Proceedings of the 1st ACM/IEEE-CS joint conference on Digital libraries*, pages 1–10.

Rehm, G. (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 Hawaii Internat. Conf. on System Sciences*.

Rehm, G., Santini, M., Mehler, A., Braslavski, P., Gleim, R., Stubbe, A., Symonenko, S., Tavosanis, M., and Vidulin, V. (2008). Towards a reference corpus of web genres for the evaluation of genre identification systems. In *Proceedings of the Sixth Language Resources and Evaluation Conference, LREC 2008*, Marrakech.

Santini, M. (2009). Cross-testing a genre classification model for the web. In Mehler et al. (2009).

Sharoff, S. (2006a). 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.

Sharoff, S. (2006b). Open-source corpora: using the net to fish for linguistic data. *International Journal of Corpus Linguistics*, 11(4):435–462.

Sharoff, S. (2009). In the garden and in the jungle. Comparing genres in the BNC and Internet. In Mehler et al. (2009).

Stamatatos, E., Kokkinakis, G., and Fakotakis, N. (2000). Automatic text categorization in terms of genre and author. *Computational Linguistics*, 26(4):471–495.

Stein, B., Meyer zu Eissen, S., and Lipka, N. (2009). Web genre analysis: Use cases, retrieval models, and implementation issues. In Mehler et al. (2009).

Stubbe, A. and Ringlstetter, C. (2007). Recognizing genres. In *Abstract Proceedings of the Colloqium "Towards a Reference Corpus of Web Genres*.

Swales, J. (1990). *Genre Analysis. English in academic and research settings*. Cambridge University Press, Cambridge.

Symonenko, S. (2007). Recognizing genre-like regularities in website content structure. In *Towards Genre-Enabled Search Engines: The Impact of Natural Language Processing*.

Toms, E. and Campbell, D. (1999). Genre as interface metaphor: exploiting form and function indigital environments. In *Proc. 32nd Annual Hawaii International Conference on System Sciences*.

Varga, D., Halacsy, P., Kornai, A., Nagy, V., Nemeth, L., and Tron, V. (2007). Parallel corpora for medium density languages. In N. Nicolov, K. Bontcheva, G. A. and Mitkov, R., editors, *Recent Advances in Natural Language Processing IV. Selected papers from RANLP-05*, pages 247–258. Benjamins.

Yates, J. and Orlikowski, W. (1992). Genres of organizational communication: A structurational approach to studying communication and media. *Academy of management review*, pages 299–326.