Riding the Rough Waves of Genre on the Web
Concepts and Research Questions

Marina Santini[1], Alexander Mehler[2], Serge Sharoff[3]

[1] HATII, University of Glasgow, UK
MarinaSantini.MS@gmail.com
[2] Faculty of Technology, Bielefeld University, Germany
Alexander.Mehler@uni-bielefeld.de
[3] Centre for Translation Studies, University of Leeds, UK
s.sharoff@leeds.ac.uk

1 Why is Genre Important?

Genre, in the most generic definition, takes the meaning "kind; sort; style" (OED). A more specialised definition of genre in OED reads: "A particular style or category of works of art; esp. a type of literary work characterised by a particular form, style, or purpose.". Similar definitions are found in other dictionaries, for instance, OALD reads "a particular type or style of literature, art, film or music that you can recognise because of its special features". Broadly speaking, then, generalising from lexicographic definitions, genre can be seen as a classificatory principle based on a number of characterising attributes.

Traditionally, it was Aristotle, in his attempt to classify existing knowledge, who started genre analysis and defined some attributes for genre classification. Aristotle sorted literary production into different *genre classes* by focussing on the attributes of purpose and conventions[4].

After him, through the centuries, numberless definitions and attributes of the genre of written documents have been provided in differing fields, including literary criticism, linguistics and library and information science. With the advent of digital media, especially in the last 15 years, the potential of genre

[4] More precisely, "in the Poetics, Aristotle writes, 'the medium being the same, and the objects [of imitation] the same, the poet may imitate by narration – in which case he can either take another personality as Homer does, or speak in his own person, unchanged – or he may present all his characters as living and moving before us' ... The Poetics sketches out the basic framework of genre; yet this framework remains loose, since Aristotle establishes genre in terms of both convention and historical observation, and defines genre in terms of both convention and purpose". Glossary available at The Chicago School of Media Theory, retrieved April 2008.

for practical applications in language technology and information technology has been vigorously emphasised by scholars, researchers and practitioners.

But why is genre important? The short answer is: because it reduces the cognitive load by triggering expectations through a number of conventions. Put in another way, genres can be seen as sets of *conventions* that transcend individual texts, and create frames of recognition governing document production, recognition and use. Conventions are *regularities* that affect information processing in a repeatable manner [29]. Regularities engage *predictions* about the "type of information" contained in the document. Predictions allow humans to identify the *communicative purposes* and the *context* underlying a document. Communicative purposes and context are two important principles of human communication and interactions. In this respect, genre is then an implicit way of providing background information and suggesting the cognitive requirements needed to *understand a text*. For instance, if we read a sequence of short questions and brief answers (*conventions*), we might surmise that we are reading FAQs (*genre*); we then realize that the purpose of the document is to instruct or inform us (*expectations*) about a particular topic or event of interest. When we are able to identify and name a genre thanks to a recurrent set of regular traits, the functions of the document and its communicative context immediately build up in our mind. Essentially, knowing the genre to which a text belongs leads to predictions concerning form, function and context of communication. All these properties together define what Bateman calls the "the most important theoretical property" of genre for empirical study, namely the power of *predictivity* [9, p.196]. The potential of predictivity is certainly highly attractive when the task is to come to terms with the overwhelming mass of information available on the web.

1.1 Zooming In: Information on the Web

The immense quantity of information on the web is the most tangible benefit (and challenge) that the new medium has endowed us as web users. This wealth of information is available either by typing a URL (suggested by other web external or web internal sources) or by typing a few keywords (the query) in a search box. The web can be seen as the *Eldorado* of information seekers.

However, if we zoom in a little and focus our attention on the most common web documents, i.e. written texts, we realize that finding the "right" information for one's need is not always straightforward. Indeed, a common complaint is that users are overwhelmed by huge amounts of data and are faced with the challenge of finding the most relevant and reliable information in a timely manner. For some queries we can get thousands of hits. Currently, commercial search engines (like Google and Yahoo!) do not provide any hint about the *type of information* contained in these documents. Web users may intuit that the documents in the result list contain a *topic* that is *relevant* to their query. But what about other dimensions of communication?

As a matter of fact, Information Retrieval (IR) research and products are currently trying to provide other dimensions. For instance, some commercial search engines provide specialised facilities, like Google Scholar or Google News. IR research is active also in plagiarism detection[5], in the identification of context of interaction and search[6], in the identification of the "sentiment" contained in a text[7], and in other aspects affecting the reliability, trust, reputation[8] and, in a word, the appropriateness of a certain document for a certain information need.

Still, there are a number of other dimensions that have been little explored on the web for retrieval tasks. *Genre* is one of these. The potential of genre to improve information seeking and reduce information overload was highlighted a long time ago by Karlgren and Cutting [47] and Kessler et al. [48]. Rosso [78] usefully lists a pros and cons of investigating web retrieval by genres. He concludes on a positive note, saying that genre "can be a powerful hook into the relevance of a document. And, as far as the ever-growing web is concerned, web searches may soon need all the hooks they can get". Similarly, Dillon [29] states "genre attributes can add significant value as navigation aids within a document, and if we were able to determine a finer grain of genre attributes than those typically employed, it might be possible to use these as guides for information seekers".

Yet, the idea that the addition of genre information could improve IR systems is still a hypothesis. The two currently available genre-enabled prototypes – X-SITE [36] and WEGA (Stein et al. in this volume) – are too preliminary to support this hypothesis uncontroversially. Without verifying this hypothesis first, it is difficult to test genre effectiveness in neighbouring fields like human-computer interaction, where the aim is to devise the best interface to aid navigation and document understanding (cf. [29]).

IR is not the only field that could thrive on the use of genre and its automatic classification. Traditionally, the importance of genre is fully acknowledged in research and practice in qualitative linguistics (e.g. [98]), academic writing (e.g. [18]) and other well-established and long-standing disciplines.

However, also empirical and computational fields – the focus of this volume – would certainly benefit from the application of the concept of genre. Many researchers in different fields have already chosen the *genre lens*, for instance in corpus-based language studies (e.g. [24]; [59]; [14]), automatic summarisation [89], information extraction [40], language learning [58], creation of language corpora [84], e-government (e.g. [37], information science (e.g. [39] or [70]), information systems [72] and many other activities.

[5]For instance, see "PAN'09: 3rd Int. PAN Workshop – 1st Competition on Plagiarism Detection" `http://www.webis.de/pan-09/`

[6]For instance, see "ECIR 2009 Workshop on Contextual Information Access, Seeking and Retrieval Evaluation" `http://www.irit.fr/CIRSE/`

[7]For instance, see "CyberEmotions" `http://www.cyberemotions.eu/`

[8]For instance, see "WI/IAT'09 Workshop on Web Personalization, Reputation and Recommender Systems" `http://www.wprrs.scitech.qut.edu.au/`

The genres used by Karlgren and Cutting [47] were those included in the Brown corpus. Kessler et al. [48] used the same corpus but were not satisfied with its genre taxonomy, and re-labelled it according to their own nomenclature. Finding the appropriate labels to name and refer to genre classes is one of the major obstacles in genre research (see Rosso and Haas; Crowston et al.; both in this volume). But, after all, the naming difficulty is very much connected with the arduousness of defining genre and characterising genre classes.

2 Trying to Grasp the Ungraspable?

Although undeniably useful, the concept of genre is fraught with problems and difficulties. Social scientists, corpus linguists, computational linguists and all the computer scientists working on empirical and computational models for genre identification are well aware that one of the major stumbling blocks is the lack of a shared definition of genre, and above all, of a shared set of attributes that uncontroversially characterise genre.

Recently, new attempts have been made to pin down the essence of genre, especially of web genre (i.e. the genre of digital documents on the web, a.k.a. cybergenre).

A useful summary on the diverse perspectives is provided by Bateman [9]. Bateman first summarises the views of the most influential genre schools — namely *Genre as social action* put forward by North American linguists and *Genre as social semiotic* supported by systemic-functional linguistics (SFL)[9] – then he points out the main requirements for a definition of genre for empirical studies:

> Fine linguistic detail is a prerequisite for fine-grained genre classification since only then do we achieve sufficient details (i) to allow predictions to be made and (ii) to reveal more genres than superficially available by inspection of folk-labelling within a given discourse community. When we turn to the even less well understood area involved in multimodal genre, a fine-grained specification employing a greater degree of linguistic sophistication and systematicity on the *kind of forms that can be used for evidence for or against the recognition of a genre category is even more important* ([9, p.196] – italics in the original)

Bateman argues that the current effort to characterise the kinds of documents found on the web is seriously handicapped by a relatively simple notion of genre that has only been extended minimally from traditional, non–multimodal conceptions. In particular, he claims that the definition of cybergenre, or web genres, in terms of `<content, form, functionality>`, taken

[9]The contraposition between these two schools from the perspective of teaching is also well described in Bruce [18], Chapter 2.

as an extension of the original tuple `<content, form>` is misleading (cf. also Karlgren in this volume). Also the dual model proposed by Askehave and Nielsen [4], which extends the notion of genre originally developed by Swales [91], is somewhat unsatisfying for Bateman. Askehave and Nielsen [4] propose a two-dimensional genre model in which the generic properties of a web page are characterised both in terms of a traditional text perspective and in terms of the medium (including navigation). They motivate this divide in the discussion of the homepage web genre. The traditional part of their model continues to rely on Swales' view of genre, in which he analyses genres at the level of purpose, moves and rhetorical strategies. The new part extends the traditional one by defining two modes that users take up in their interaction with new media documents: users may adopt either a reading mode or a navigation mode. Askehave and Nielsen argue that hyperlinks and their use constitute an essential extension brought about by the medium. Against this and all the stances underpinning hypertext and hyperlinking facilities as the crucial novelty, Bateman argues that the consideration that a more appropriate definition of genre should not open up a divide between digital and non digital artefacts.

Other authors, outside the multimodal perspective underpinned by Bateman [9], propose other views. Some recent genre conceptions are summarised in the following paragraphs.

Bruce [18] builds upon some of the text types proposed by Biber [11] and Biber [12] to show the effectiveness of his own genre model. Bruce proposes a two-layered model and introduces two benchmark terms: social genres and cognitive genre. Social genres refer to "socially recognised constructs according to which whole texts are classified in terms of their overall social purpose", for instance personal letters, novels and academic articles. Cognitive genres (a.k.a. text types by some authors) refer to classification terms like narrative, expository, descriptive, argumentative or instructional, and represent rhetorical purposes. Bruce points out that cognitive genres and social genres are characterised by different kinds of features. His dual model, originally devised for teaching academic writing, can be successfully applied to web genre analysis, as shown by Bruce's chapter in this volume.

The genre model introduced by Heyd [43] has been devised to assess whether email hoaxes (EH) are a case of digital genre. Heyd provides a flexible framework that can accommodate for discourse phenomena of all kinds and shapes. The author suggests that the concept of genre must be seen according to four different parameters. The vertical view (parameter 1) provides levels of descriptions of increasing specificity, that start from the most general level, passing through an intermediate level, down to a sublevel. This view comes from prototype theory and appears to be highly applicable to genre theory (cf. also [53]), with the intermediate level of genre descriptions being the most salient one. The horizontal view (parameter 2) accounts for genre ecologies, where it is the interrelatedness and interdependence of genre that is emphasised. The ontological status (parameter 3) concerns the conceptual

framework governing how genre labels should be ascribed, i.e. by a top-down or a bottom-up approach. In the top-down approach, it is assumed that the genre status depends upon the identification of manifest and salient features, be they formal or functional (such a perspective is adopted also by Sharoff in this volume); by contrast a bottom up approach assumes that the genre status is given by how discourse communities perceive a discourse phenomenon to be a genre (see Rosso and Haas; Crowston et al.; both in this volume). The issue of genre evolution (parameter 4) relates to the fast-paced advent and evolution of language on the Internet and to the interrelation with socio-technical factors, that give rise to genre creation, genre change and genre migration. Interestingly, Heyd suggests that the frequently evoked hybridity of Computer Mediated Communication (CMC) genres can be accounted for by the "transmedial stability that predominates on the functional sublevel while genre evolution occurs on the formal sublevel: this explains the copresence of old and new in many digital genres" [43, p. 201].

Martin and Rose [61] focus on the relations among five major families of genres (stories, histories, reports, explanations and procedures) using a range of descriptive tools and theoretical developments. Genre for Martin and Rose is placed within the systemic functional model (SFL). They analyse the relationship between genres in terms of a multidimensional system of oppositions related to the function of communication, e.g. instructing vs. informing.

This overview on the most recent work on genre and web genre shows that the debate on genre is still thrilling and heated. It is indeed an intellectually stimulating discussion, but do we need so much theory for a definition of web genre for empirical studies and computational applications?

2.1 In Quest of a Definition of Web Genre for Empirical Studies and Computational Applications

Päivärinta et al. [72] condense in a nutshell the view on genre for information systems:

> [...] genres arguably emerge as fluid and contextual socio-organisational analytical units along with the adoption of new communication media. On the other hand, more stabilised genre forms can be considered sufficiently generic to study global challenges related to the uses of communications technology or objective enough to be used as a means for automatic information seeking and retrieval from the web.

Essentially, an interpretation of this statement would encourage the separation of the theoretical side from the practical side of genre studies. After all, on the empirical and computational side, we need very little. Say that, pragmatically, genre represents a *type of writing, which has certain features that all the members of that genre should share*. In practical terms, and more specifically for automatic genre classification, this simply means:

1. take a number of documents belonging to different genres;
2. identify and extract the features that are shared within each type;
3. feed a machine learning classifier to output a mathematical model that can be applied to unclassified documents.

The problem with this approach is that without a theoretical definition and characterisation underpinning the concept of genre, it is not clear how to select the members belonging to a genre class and in which way the genre labels "represent" a selected genre class. A particular genre has conventions, but they are not fixed or static. Genre conventions unfold along a continuum that ranges from weak to strong genre conformism. Additionally, documents often cross genre boundaries and draw on a number of characteristics coming from different genres. Spontaneous questions then arise, including:

A) Which are the features that we want use to draw the similarities or differences between genre classes? B) Who decides the features? C) How many features are really the core features of a genre class? D) Who decides how many raters must agree on the same core feature set and on the same genre names in order for a document to belong to a specific genre? E) Are the features that are meaningful for humans equally meaningful for a computational/empirical model? And similarly F) Are genre classes that are meaningful for humans equally meaningful for a computational model? And so on and so forth.

Apparently, theoretical/practical definitions of genres have no consequence whatsoever when deciding about the *actual typification* of the genre classes and genre labels required to build empirical and computational models. This gap between definitions and empirical/classification studies has been pointed out by Andersen, who notes that freezing or isolating genre, statistically or automatically, dismantles action and context (Andersen, personal communication; cf. also Andersen [2] and Andersen [3]), the driving forces of genre formation and use. In this way, genres become *lifeless* texts, merely characterized by formal structural features.

In summary, we are currently in a situation where there is the need to exploit the *predictability* inherent in the concept of genre for empirical and computational models, while genre researchers are striving to find an adequate definition of genre that can be agreed upon and shared by a large community. Currently, the main difficulty is to work out optimal methods to define, select and populate the constellation of genres that one wishes to analyse or identify without hindering replication and comparison.

3 Empirical and Computational Approaches to Genre: Open Issues

Before moving on to the actual chapters, the next three sections focus on the most important open issues that characterise current empirical and computational genre research. These open issues concern the nature of web documents

(Section 3.1), the construction and use of corpora collected from the web (Section 3.2) and the design of computational models (Section 3.3).

3.1 Web Documents

While paper genres tend to be more stable and controlled given the restrictions or guidelines enforced by publishers or editors, on the web centrifugal forces are at work. Optimistically, Yates and Sumner [99] and Rehm [77] state that the process of imitation and the urge for mutual understanding act as centripetal forces. Yet, web documents appear much more uncontrolled and unpredictable if compared to publications on paper.

First of all, what is a web document? On the web, the boundary of a document is unclear. Is a web document a single file? If so, a frame composing a web page could be an autonomous web document. Or is it the individual web page? But then where is the core information in a web page? Can we identify it clearly? Web pages can be just navigational or both navigational and content bearing. How many autonomous texts can be found in a individual web pages? Maybe it is safer to identify the web document with a web site as a whole? Where then is the boundary of a web site?

It appears evident that on the web the granularity of documents cannot be kept implicit, because texts with different content and functions are tiled and connected together more tightly than on paper documents, where the physical pages act, sometimes, as "fences" that separate different contents and functions.

For instance, if we compare a daily newspaper like *The Times*, and its web counterpart, *Timesonline*[10], we can realize that the "paper" gives a much more static status to the concept of "document". On the paper too, a document can be interpreted at various degrees of granularity. For instance, a single text (like an editorial or a commercial advertisement) is a document; a page (like the newspaper frontpage) is a document; and a medium (like a newspaper or a book) is a document as well. But on the web, hyperlinking, search facilities, special features (like dynamic marquees), and other technicalities make the concept of documents much more dynamic and flexible. This is evident if we compare the same document granularity on the paper and on the web. Figure 1 shows an online frontpage (LHS) and a paper frontpage (RHS). Both the graphic appearance and the functionality associated with these documents differ. The basic idea of providing an entry point with snippets of the contents is maintained in both media[11], but the online frontpage has also a corollary of interactive activities, such as menus, search boxes, and dynamic texts. Additionally, past editions or news articles are immediately available by clicking on the archive link. While the paper frontpage is

[10] Global edition: `http://www.timesonline.co.uk/tol/global/`, or UK edition `http://www.timesonline.co.uk/tol/news/`

[11] As noted by Bateman [9] functionality belongs to both paper and web documents.

Fig. 1. Frontpage of a web newspaper vs. its printed counterpart.

a self-contained unity, with internal cross-references and occasional citations to external sources, the online frontpage has no boundaries, each web page or each section of a web page can be connected to both internal and external pages. Interactivity, multimodality and dynamic content make the online frontpage different from a paper frontpage. While the paper frontpage has the physical boundary of the first page in a newspaper, and one can dwell on it, the online frontpage is a gateway, i.e. a navigational page providing access to other pages. It becomes clear, then, that when working with web documents, while all levels of granularity are plausible, there is the need to spell out explicitly and justify the *unit of analysis*.

Essentially, web genres are composite functional types of web-based communication. For this reason, in order to make them an object of automatic classification we need to decide on the reference units of their manifestations. That is, we need to decide which document structures of the web are attributed to web genres: e.g., self-contained pages [80] or their constituents [76, 90, 77, 96], websites [66, 57] or even larger units such as, for example, domains consisting of several websites [15]. When it comes to modelling such web document structures as instances of web genres, we realise that the vector space approach (see Part III of this volume) is only one of many ways to model genre computationally. One reason is that if one had to choose a single characteristic of genres on the web, then the linkage of their instances by hyperlinks would be a prime candidate (see Part IV of this volume). Web genres are manifested by pages [80, 81] that are interlinked to create, in effect, larger units *above the level of single pages*. Thus, any decision on the manifestation unit of web genres should clarify the role of hyperlink-based structure formation as a source of attributing these units to the focal web genres.

With respect to web content mining, Menczer [69] observes that the content of a page is similar to that of the pages that link to it. We may vary this

link-content conjecture by saying that you shall know a web genre (though not solely) by the link-based neighbourhood of its instances. Following this line of thinking we can distinguish three levels of modelling web documents as instances of web genres (cf. [63, 77]):

- On the *micro level* we analyse page-level [79] units and their constituents [90] as self-contained (though not necessarily the smallest) manifestations of web genres. These then enter into websites as more complex web genre units.
- On the *meso level* we deal with single or conglomerate websites and their web-specific structure formation which, of course, is hardly found beyond the web [15].
- On the *macro level* we deal with the web as a whole from the perspective of complex network analysis and related approaches [30].

In order to exemplify the differences of these three perspectives, take social software as an example: here, web genre analysis may focus microscopically on single weblogs [71] as instances of this genuine web genre or on networks of blogs which are interlinked by trackbacks and related means [42, 51]. From the point of view of a mesoscopic perspective we may analyse, more specifically, *blog sites* as sub-networks of networked blogs whose connection may result from their discussion of a common topic [51]. Last but not least, we gain a macroscopic perspective by taking into account blog network-external links which embed blogs into the web as a whole. Analogously, by analysing Wikipedia as an instance of web-based knowledge communication we may distinguish wiki-internal structures (e.g. in the form of portals) from wiki-external structures (by analysing links from wikis to pages of external sites) [62].

Genre research has focussed mostly on analysing micro and meso level units as instances of web genres (see, for example, the contributions of Björneborn [16] and Santini [82]). One might hesitate to consider macro level approaches under this perspective. However, by analogy to text genres we know of the existence of macro genres which are generated from instances of different (micro-level) genres [60]. In the web, this build-up of macro genres is more explicit on the instance level as authors make use of hyperlinks to interconnect micro or meso level units of the same macro genre. Further, the macro-level perspective opens the chance to study both the network of web genres as a network of hypertext types (which evolve as part of the same semiotic universe) as well as the network of their instances. This gives a bipartite perspective on networking on the level of hypertext types and their instances which is nearly inaccessible to text genre analysis.

Björneborn [15] (and in this volume) offers a rich terminology by distinguishing four nested levels of structure formation (i.e., pages, directories, domains and sites) together with a typology for the perspective classification of a link. A university website, for example, is described as comprising different websites of various genres (among other things, the difference between project

homepages and personal academic homepages) whereas, together with other university websites, it forms the domain of academia. Thelwall et al. [94] generalise this model in terms of the *Alternative Document Model*. They do that by additionally distinguishing *web spaces* as sub-networks of web documents demarcated, e.g., by geographic criteria.

If we, on the other hand, look on the micro level of structure formation in the web, we see that the notion of *logical document structure* dominates the corresponding range of models. By analogy to text documents [74] the idea is that the attribution of a web document to a web genre is made more difficult by insufficiently explicit logical document structures. This can come as a result of, e.g., the abuse of tags [6] or the failure to use hyperlinks to connect functionally homogeneous, monomorphic document units [67]. Manifestations of webgenres are analysed, for example, as *compound documents* [31], as *logical domains* [54], as *logical documents* [92, 55] or as *multipage segments* [25].[12] Whatever is seen to be the exact unit of manifestation of a web genre — say on the page level, below or above — approaches to learning corresponding classifiers face the formation of hyperlink-based, network-inducing structures apart from purely hierarchical text structures. Notwithstanding these differences we have to state that whatever is seen to be the exact unit of manifestation of a web genre — say on the page level, below or above — the corresponding classifiers, in their approach to learning, face the challenge of forming hyperlink-based, network-inducing structures that are fundamentally different from [or more complex than] purely hierarchical text structures. We suggest that more complex graph models (above the level of tree-like structures) are needed to bring into focus the web genre modelling of the future, which complete and complement the more traditional vector space approaches.

One obvious consequence of the composite and diversified characterisation of web documents is the necessity to devise *classification schemes* not constrained to the single genre class assignment. Intuitively, there is a high likelihood that many web documents (whatever their granularity) would fall into multiple genre classes, and many would remain unclassified by genre because of a high degree of individualisation or hybridism. Genre analysts also point out that the acknowledgement and usage of genres are subjective and depend upon membership in a discourse community (cf. Crowston et al. in this volume). The flexibility of a classification scheme would then account also for the subjectivity of use and recognition of genres by web users. Since the web serves many communities and web users are exposed to innumerable contacts, it would be wiser to devise a classification scheme addressing this complexity in the future.

Importantly, the nature and the unit of analysis of web documents has not only repercussions on genre classification schemes, but also affects *genre evolution*. Genres are historical entities, they develop over time, and in response

[12] See also Tajima et al. [93], Cohn and Hofmann [23] and Chakrabarti et al. [22] for topic-related approaches in this line of research.

to social, cultural and technological contexts (e.g. see Paolillo et al. in this volume). Existing genres may simply go out of fashion, or undergo transformation. Frequently, genres on the web evolve when they migrate from one medium to another (see the frontpage example above). They can also be created from scratch, due to new web technologies or new contexts of interaction. The personal home page and blog genres are the classical examples of web genres whose existence cannot be imagined outside the web. The formation of new genres from an antecedent can also be monitored computationally [65]. For example, it is easily predictable that the recent booming of social networks – from Facebook to Twitter and LinkedIn – will presumably destabilise and change web genres like the personal home page and blog that were thought to be "novel" just some months ago. The technology offered by social networks in creating personal profiles, live feeds, blogging, notes and material of any kind at the same time are clear signs that new genres are going to materialise soon.

In summary, web documents would require a flexible genre classification scheme capable of making sense of 1) the composite structure of web documents at any level of unit of analysis; 2) the complexity of interaction allowed by web documents; 3) the subjective and differing naming conventions due the membership to different communities and finally 4) the tendency towards rapid change and evolution of genre patterns.

3.2 Corpora, genres and the web

According to John Sinclair, a corpus is "a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language" [87]. Criteria for selecting texts for a corpus can include information about the authorship, audience or domain of its constituent texts, but selection of texts by their genre is nearly always present as one of the main criteria for designing a traditional corpus. For instance, the Brown Corpus, the first computer corpus developed in the 1960s, was compiled using the following linguistic criteria [52]:

- it was restricted to texts written originally in English by native speakers of American English (as far as this can be determined);
- the texts were first published in the United States in 1961;
- samples of entire texts were selected starting from a random sentence boundary and ending by the first sentence boundary after an uninterrupted stretch of 2000 words (this means that texts themselves had to be longer than 2000 words);
- texts were selected from 15 text categories: A) Press: reportage, B) Press: editorial, C) Press: Reviews, D) Religion, E) Skill and hobbies, F) Popular lore, G) Belles-lettres (biography, memoirs, etc.), H) Miscellaneous: US Government & House Organs, J) Learned (i.e., research articles), K) Fiction: general, L) Fiction: mystery and crime, M) Fiction: science, N)

Fiction: adventure and western, P) Fiction: romance and love story, R) Humor.

As we can see from this specification, the only variation among samples present in the Brown Corpus concerns their text categories, which roughly correspond to genres (the only possible exceptions are Religion and Skills&Hobbies, but even they constitute distinct functional styles, which are normally associated with specific genres, i.e., sermons and DIY magazines).

Further development of corpora, e.g., creation of the Bank of English [86], the British National Corpus [5], or the American National Corpus [44], resulted in a greater variety of parameters for describing their constituent texts, but they nevertheless classified them into genres, even if the genres in each corpus were defined in various incompatible ways. For instance, the original release of the BNC classified the written texts into their publication medium (e.g., book or periodical), domain (commerce, social sciences or imaginative), and target audience. This provided an opportunity to specify some genres by restricting one or more BNC metadata tags, e.g., fiction corresponds to imaginative texts, research papers can be found by a combination of tags coding texts from natural, applied or social sciences, aimed at the professional audience, and not published as books. Since this situation was treated as less than adequate, David Lee developed a system of 70 genre tags for BNC documents [53], e.g., `W_ac_natsci` or `W_ac_socsci` for academic papers in the domains of natural or social sciences.[13]

The situation with genres in web-derived corpora is a bit different. The majority of large web corpora have not been collected in any pre-planned way with respect to their target domains or genres. Collection of texts from the web normally involves taking publicly accessible documents from a list of URLs. This means it is driven by the availability of sources, which leaves many parameters of corpus collection, such as genres, unspecified.

Some web corpora are created by "focused crawling", which, in its simplest form, involves selecting several websites containing a large number of texts which are of interest to the corpus collector, and retrieving the entire set of texts from these websites, e.g., the entire Wikipedia or webpages of major universities. More advanced methods of focused crawling involve starting with a seed set of links and then collecting links to other relevant websites, with the relevance assessed by keywords and/or hypertext links between pages, as similar pages tend to have more inter-connections with each other [21]. In all cases of focused crawling, the seed set of URLs used for collecting a web corpus restricts its range of genres, but does not define it precisely. For instance, articles retrieved from Wikipedia can be biographies, time-lines of events, introductions to academic theories, some subtypes of news items, etc., but they cannot include such genres as blogs, fiction, humour or memoirs.

[13] This is another example where a difference in the domain of a text contributes to a difference in its genre.

Another method for corpus collection relies on making automated queries to a major search engine and retrieving webpages for the top N (10-20-100) URLs returned by it. The choice of keywords affects the composition of the resulting corpus to some extent. For instance, if a large number of specialised terms are used in queries, e.g., *amnesia, myoclonic, paroxysmal*, the resulting corpus will contain mostly highly technical medical texts and relatively few patient leaflets or news items. Using common words from the general lexicon, e.g., *picture, extent, raised, events*, results in a corpus with a variety of domains and text types [83]. On the other hand, queries using function words (*the, of, to*) result in a larger number of index pages [34].

Finally, web corpora usually contain a very large number of relatively small documents. The Brown Corpus contains 500 documents. The BNC, being 100 times bigger in terms of word count, contains just 4055 distinct documents, many of which are composite texts collected from entire issues of newspapers, journals or radio programmes. Given a small number of texts in traditional corpora it was feasible to annotate them with respect to genres while they were collected. On the other hand, the number of documents in web corpora is considerably larger, e.g., exceeding two million webpages for Web-as-Corpus projects developed at the University of Bologna [7, 33]. Thus, their manual annotation is practically impossible. Their genre composition is usually assessed indirectly by studying samples of their texts or by comparing the frequencies of keywords extracted from them (however, see Part III of this volume for a variety of methods for automatic classification of texts by genre).

There are at least three factors that can influence the distribution of genres in web-derived corpora:

- some genres are not well represented on the web;
- a large number of documents are located in the "hidden web", which is not accessible to crawling;
- the process of corpus collection usually puts restrictions on file types retrieved from the web.

The web is an enormous resource, with more and more texts appearing there in a variety of languages. However, many genres are still underrepresented. This primarily concerns copyrighted work aimed at a wider public audience, such as fiction and non-fiction recreational reading. Their authors expect to receive royalties for their effort, and their publishers do not normally provide free online access. Texts in these genres do appear on the web, for instance, many amateur science-fiction authors regular publish their works electronically under a Creative Commons licence, and Project Gutenberg collects out-of-copyright fiction. However, the selection available on the web is significantly skewed in comparison to offline fiction.

The hidden web (also called Deep Web) consists of pages that are difficult to access by crawling. Some of them are dynamically generated in response to a user query, e.g., some archived news items are stored in a database and can be retrieved only by specifying their date or keywords. Some hidden webpages

are ordinary webpages which are not linked to any visible webpage, or which are accessible only by a password (not usually available to the crawler) or via a mechanism requiring some kind of user interaction, e.g., Javascript-based selection. Some estimates put the total size of the hidden web to be 500 times bigger than the surface web accessible to major search engines [41]. The hidden web is particularly important for search engines, as their aim is to index every possible webpage. This concern is less important for corpus collection, as a corpus is only a sample of the totality of texts in a given language. However, understanding the composition of the hidden web is important as it affects the distribution of genres. For instance, short descriptions of a large number of resources, such as synopses of books in a library, are more likely to be in the hidden web (accessible by queries to book names), so they are more likely to be underrepresented in web-derived corpora.

Finally, some file types are inherently easier to deal with. For instance, it is easy to retrieve plain text content from HTML pages, so HTML pages are more often used for corpus collection in comparison to, say, Word documents, which need special tools for retrieving textual content. PDF and Postscript files are commonly used on the web to present publishable information, such as books, articles or brochures. However, in terms of their internal format they contain a sequence of drawing primitives, often, but not necessarily, corresponding to characters, so that it is difficult to reconstruct the flow of text, spaces between words or even the encoding of non-Latin characters. The situation with Flash objects (normally containing animation, but often presenting a large amount of text) is even worse, as their drawing primitives include motion of respective objects across the computer screen. In the end, many formats apart from plain HTML files are often omitted from web-derived corpora, skewing their genre diversity. In the modern web this is especially important for PDF files, which are the preferred format for final typeset products, such as catalogues, published research results or white papers. Often these texts are not available in the form of HTML files.

In summary, although web corpora are designed to contain examples of texts in exactly the same way as traditional corpora are, they are different in some respects and there is no consensus on many important aspects.

In addition to the construction issues outlined above, there are also other controversial issues related to formatting and cleaning webcorpora. In many cases traditional corpora were produced by scanning hard copies of texts and applying OCR (optical character recognition) to the result. In other cases, texts were typed in from scratch. In either case, traditional corpora do not preserve much information about formatting, with the only possible exception of paragraph boundaries. In the end, a text stored in a traditional corpus often consists of a flat sequence of sentences with little typographic information preserved.[14]

[14] After collecting texts, developers of traditional corpora often introduce their own set of annotation layers, such as POS tagging, semantic or metatextual markup,

On the other hand, Web corpora coming from HTML pages contain relatively rich markup. As far as corpus collection is concerned, this markup takes three different forms:

1. navigation frames enabling navigation on a complex website (topics/subtopics, pages on related topics, calendar links, etc); and
2. text-internal hyperlinks, when running text is enriched with hypertextual markup linking to other relevant documents or other sections of the same document;
3. non-hypertextual markup, such as explicit formatting of headings, lists, tables, etc.

When webpages are collected to be used as a corpus for linguistic studies, one approach to corpus collection pays more attention to selecting running text. In this approach extra efforts are devoted to cleaning webpages from unwanted navigation frames [8]. The rationale behind this 'cleaning' approach is to make web-derived corpora useful for research in natural language processing, lexicography or translation, because expressions frequently occurring in navigation frames, such as *Current events*, *See also* or *Have your say*, can considerably distort the language model. Similarly, text-internal links are often discarded, while their text remains, so that web corpora become more similar to their traditional counterparts.

Some portions of non-hypertextual markup in the form of headings and lists are often preserved in the cleaning approach, since deletion of this information again distorts the language model by introducing incomplete sentences within standard running text. Finally, some markup present in many webpages is used for presentational purposes only. For example, web designers often introduce table cells to separate different parts of text, e.g., navigation frames from the main body, or a new reply message in a forum from a quote from a previous message, whereas from the viewpoint of the content, such elements can be considered as distinct paragraphs. Therefore, the cleaning approach normally discards information about tables or replaces them with paragraph boundaries.

This approach to collecting and distributing webcorpora is useful in some respects, since it makes web-derived corpora closer to their offline counterparts. However, it discards a lot of information and makes the study of unique features of web genres more difficult. This also makes it harder to detect web genres automatically, as some crucial information for genre detection is present in the form of discarded features, e.g., navigation frames are more common in particular genres, and, similarly, documents of the same genre are often cross-linked. As a matter of fact, many genre collections built for classification purposes maintain original webpages in their entirety without attempting to clean them artificially (e.g. see the KI-04 corpus and

---

but such layers are not taken from original texts in the form they have been published.

the 7-webgenre collections described in Santini, in this volume; see also the super-genre collection used by Lindemann and Littig, in this volume).

In summary, at the current stage of genre research no standards have been agreed for the construction of web genre corpora. Decisions, choices and operationalisations are made subjectively, following individual needs. However, projects are put forward to established shared standards (see "Any Land in Sight", the concluding chapter of this volume).

3.3 Empirical and Computational Models of Web Genres

The approach dominating automatic genre identification research is based on supervised machine learning, where each document is represented like a vector of features (a.k.a. the vector space approach), and a supervised algorithm (e.g. SVM or NB) automatically builds a genre classification model by "learning" from how a set of features "behave" in exemplar documents (e.g. see Sharoff; Kim and Ross; both in this volume). Many different feature sets have been tried out to date, e.g. function words, character n-grams, Parts of Speech (POS), POS tri-grams, Bag of Words (BOW), or syntactic chunks. Most of these feature sets have been tested on different genre corpora, differing in terms of number and nature of genres, and in terms of number of documents per genre. Although some comparative experiments have been carried out, the absence of genre benchmarks or reference corpora built with shared and agreed upon standards makes any comparison difficult, because existing genre collections have been built with subjective criteria, as pointed out in the previous section. A partial and temporary remedy to this situation has been adopted recently, i.e. cross-testing (see Santini, in this volume).

Although the vector space approach is, for the time being, the most popular approach, in this last section of the open issues, we would like to outline a more complex view of web genres as source of inspiration and food for thought in future research. In Section 3.1, we suggested locating instances of web genres on, above and below the level of websites. The decision on this *manifestation level* belongs to a series of related decisions which have to be made when it comes to modelling web genres. In this section, we briefly describe four of these decisions when the focus is on *structure*.

- *Deciding on the level of web genre units as the output objects of web genre classification:* Lindemann and Littig (in this volume) present a model of web genre classification at what they call the *supergenre* level. This concerns a level of functional units which are composed of one or more genre level units. Interestingly, Lindemann and Littig consider websites as manifestation units of these supergenres. From that perspective we get the level of supergenres, of genres themselves and of subgenres as candidate output objects of a web-genre-related classification. Note that we may alternatively speak of macro, meso and micro (level) genres as has been done above. Conversely, Santini (in this volume) and all approaches

reviewed by her consider generic units of a comparative level of abstractness, but focus on web pages as their manifestation units. This divergence opens the possibility of a many-to-many relation between the output units of classification, i.e., the types which are attributed, and the input objects of classification, that is, the instances to which these types are attributed. Thus, by opting for some micro-, meso- or macro-level web genres one does not automatically determine the manifestation unit in the form of websites, web pages or page constituents. From that perspective, a decision space is created in which any location should be substantiated to keep replicability of the model and comparability with related approaches. By looking for what has been done towards such a systematisation we have to state that it is like weeding the garden, and that we are rather at the beginning.

- *Deciding on the level of manifestation units as the input objects of web genre classification:* the spectrum of this decision has already been outlined above.
- *Deciding on the features to be extracted from the input objects as reference values of classification:* when classifying input objects (e.g. web pages or sites) by attributing them to some output units (as elements of a certain genre palette), we need to explore certain features of the input objects. Among other things, we may explore *distinctive features* on the level of graphemes [46, 57], *linguistic features* in a more traditional sense [17, 38, 49, 82, 85, 88], *features related to non-hyperlink-based discourse structures* [19] or *structural features induced by hyperlinks* [16, 26, 57, 65]. In Section 3.1 we put special emphasis on less-frequently considered structure-related features of web genres. This is done according to the insight that they relate to an outstanding characteristic of genres on the *web*.
- *Deciding on the classifier model to be used to perform the classification:* facing complementary or even competing feature models as being inevitable in web genre modelling, composite classifiers which explore divergent feature resources have been common in web genre modelling from the beginning [45]. In line with this reasoning we may think of web genre models which simultaneously operate on nested levels of generic resolution. More specifically, we may distinguish *single-level* from *multi-level* approaches, which capture at least two levels of web genre structuring: that is, approaches which attribute, for example, genre categories to websites subject to attributing subgenre categories to their elementary pages (other ways of defining *two-level* genre models can be found in Santini; Bruce; both in this volume) . Note that the majority of approaches to web genre modelling realize single-level models by mapping web pages onto genre labels subject to one or more bag-of-features models. For this reason, multi-level approaches may be a starting point for building future models in this area.

By analogy to Biber [13] we may say that the structure of a web document correlates with its function, that is, with the genre it manifests. In other words: different genres have different functions, so that their instances are structured differently. As a consequence, the structure of a web document, whether a site, page or page segment, can be made a resource of feature extraction in web genre tagging. We summarise five approaches focussing on structure in the following list:

- *Bag-of-Structural-Features Approaches:*
  A classic approach to using structural features in hypertext categorisation is from Amitay et al. [1] — see Pirolli et al. [73] for an earlier approach in this line of research. Amongst others, Amitay et al. distinguish up, down, side and external links by exploring directory structures as manifested by URLs. They then count their frequencies as structure-related features. The idea is to arrive at a bag-of-structural features: that is, to analyse reference units whose frequencies are evaluated as dimensions of corresponding feature vectors. A comprehensive approach to using structure-related features in line with this approach is proposed by Lindemann and Littig [57].[15] They explore a wide range of features, similar to Amitay et al. [1], by including features which, amongst others, are based on the file format and the composition of the URL of the input pages. See also Kanaris and Stamatatos [46] who build a *bag of HTML tags* as one feature model of web genre classification (see Santini [82] for a comparative study of this and related approaches).
  Generally speaking, linguistics has clarified the fundamental difference between explicit layout structure, implicit logical (document) structure and hidden semantic or functional structure [13, 10, 74]. From that perspective one does not assume, for example, that URL-based features are reliable indicators of logical web document structures. Rather, one has to assume — as is done by Lindemann and Littig [57] — an additional level of the manifestation of web genres, that is, their *physical storage* (including file format and directory structures). In any event, it is important to keep these structural levels apart as these are different resources for guessing the functional identity of a website. This can be exemplified by Amitay et al. [1] who introduce the notion of a *side link*, which exists between pages located in the same directory (cf. Eiron and McCurley [31] for a directory-based notion of up, down and side links). It is easy to construct an example where a side link, which in terms of its physical storage manifests a *paratactic* link, is actually a *hypotactic* down or up link when being considered from the point of view of logical document structure [63]. Thus, any approach which explores structural features should clarify its standpoint regarding the difference of physical storage, layout and logical document structure.

[15] See Lim et al. [56] for a study of the impact of different types of features including structural ones.

- *Website-Tree- and Page-DOM-related Models:*
  A bag-of-structural-features approach straightforwardly adapts the bag-of-words approach of text categorisation by exploring the link and page structure of a site. This is an efficient and easy way to take web structure into account [57]. However, a more expressive and less abstract way to map this structure is to focus on the hierarchical *Document Object Model* (DOM) of the HTML representation of pages [28] or, additionally, on the mostly hierarchical kernel of the structure of a website [32]. Starting from the tree-like representation of a website, Ester et al. [32] build a Markov tree model which predicts the web genre C of a site according to the probability that the paths of this tree have been generated under the regime of C. Tian et al. [95] build a related model based on a hierarchical graph model in which the tree-like representation of websites consists of vertices which denote the DOM tree of their elementary pages. See Diligenti et al. [28], Frasconi et al. [35] and Raiko et al. [75] for related models of web document structures. See Chakrabarti [20] for an early model which explores DOM structure for hypertext categorisation (however with a focus on topical categorisations). Further, see Wisniewski et al. [97] for an approach to transforming DOM trees into semantically interpreted document models.
- *Beyond Hierarchical Document Models:*
  The preceding paragraph has presented approaches which start from tree-like models of web documents. This raises the question for approaches based on more expressive graph models. Such an alternative is proposed by Dehmer and Emmert-Streib [26]. Their basic idea is to use the page or site internal link structure to induce a so-called *generalised tree* from the kernel document structure, say, a DOM tree. The former is more informative than the latter as it additionally comprises up, down and lateral edges [64] which generalise the kernel tree into a graph. Note that this approach is powerful enough to represent page internal and external structures and, therefore, grasps a large amount of website structure. However, it maps structured data onto feature vectors which are input to classical approaches of vector-based classifications and, thus, departs from the track of Markov modelling. See Denoyer and Gallinari [27] who develop a Markov-related classifier of web document structures which, in principle, can handle *Directed Acyclic Graph*s (DAG). See alternatively Mehler [65] who develops a structure-based classifier of social ontologies as part of the Wikipedia. Extending the notion of a generalised tree, this model generalises the notion of a DAG in terms of *generalised nearly acyclic directed graph*s in order to get highly condensed representations of web-based ontologies with hundreds and thousands of vertices.
- *Two-level Approaches to Exploring Web Genre Structures:*
  The majority of approaches considered so far have been concerned with classifying units of web documents of a homogeneous nature — whether pages, their segments or complete websites. This leaves plenty of room for considering approaches which perform, say, a generic categorisation

of websites, subject to the categorisation of their elementary pages. Alternatively, we may proceed according to a feature-vector approach by representing a website by a composite vector as the result of aggregating the feature vectors of its pages (cf. the "superpage" approach of Ester et al. 32). However, such an approach disregards the structure of a site because it represents it, once more, as a bag of features. Therefore, alternative models are required. Such an approach has been proposed by Kriegel and Schubert [50] with respect to topic-related classifications. They represent websites as vectors whose dimensions represent the topics of their pages so that the sites are classified subject to the classification of the pages. Mehler et al. [67] have shown that web genres may be manifested by whole sites, single pages or page segments. Facing this variety, the genre-related segmentation of pages and their fusion into units of the logical web document structure is an important step to grasping macro, meso and micro level units of web genres in a single model. Such a segmentation and fusion algorithm is proposed by Waltinger et al. [96] for web pages. The idea is to arrive at *monomorphic* segments as manifestations of generic units on the sub-genre level. This is done by segmenting pages using their visual depiction — as a byproduct this overcomes the tag abuse problem [6] which results from using HTML tags for manifesting layout as well as logical document structures. A paradigmatic approach to a two-level website classification which combines the multi-level manifestation perspective with a tree-like structure model is proposed by Tian et al. [95] (see above) who build a hierarchical graph whose vertices represent the DOM structure of the page constituents of the corresponding site.

- *Multi-Resource Approaches — Integrating Thematic with Structural Features:*
  Almost all approaches discussed so far focus on structural features. However, it is obvious that one must combine structural with content-related features by considering the structural position of content units within the input pages. See, for example, Joachims et al. [45] who study combined kernels trained on bag-of-words and bag-of-links models, respectively. See also Tian et al. [95] who integrate a topic model with a DOM-related classifier — however with a focus on thematic classification.

In summary, as already suggested in Section 3.1, more focus on structure is needed to enhance web genre modelling in the future. We conjecture that a closer interaction between vector space approaches and structure-oriented methods can increase our understanding of web genres as a whole, thus providing a more realistic computational representation of genres on the web.

4 Conclusions

In this introduction, we emphasised why the study of genres on the web is important, and how empirical studies and computational models of web genres, with all their challenges, are the cutting edge of many research fields.

In our view, modern genre research is no longer confined to philosophical, literary and linguistic studies, although it can receive enlightment from these disciplines. Undoubtedly, Aristotle, with his systematic classificatory mind, can still be considered the unquestioned initiator of genre studies in the Western World [16]. However, modern genre research transcends the manual and qualitative classification of texts on paper to become a meta-discipline that contributes to and delves into all the fields grounded in digital media, where quantitative studies of language, language technology, information and classification systems, as well as social sciences play an important role.

In this respect, this volume contributes to the current genre discussion in six ways:

1. It depicts the state of the art of genre research, presenting a wide range of conceptualisations of genre together with the most recent empirical findings.
2. It presents an overview of computational approaches to genre classification with a special emphasis on structural models.
3. It focuses on the notion of genres "for the web", i.e., for the medium that is pervading all aspects of modern life.
4. It provides in-depth studies of several divergent genres on the web.
5. It points out several representational, computational and text-technological issues that are specific to the analysis of web documents.
6. Last but not least, it presents a number of intellectually challenging positions and approaches that, we hope, will stimulate and fertilise future genre research.

5 Outline of the Volume

Apart from the introduction, the volume is divided into four parts, each focussing on a specific facet of genre research.

PART II (*Identifying the Sources of Web Genres*) includes three chapters that analyse the selection and palettes of web genres from different perspectives.

Karlgren stresses how genre classes are both sociological constructs and stylostatistically observable objects, and how these two views can inform each

[16] There are indeed many other scholars in other parts of the world, such as the Mao school in ancient China, `http://en.wikipedia.org/wiki/Shijing`, who have pondered about the concept of genre.

other. He monitors genre variation and change by observing reader and author behaviour.

Crowston and co-workers report on a study to develop a `bottom-up' genre taxonomy. They collect a total of 767 (then reduced to 298) genre terms from 52 respondents (teachers, journalists and engineers) engaged in natural use of the Web.

Rosso and Haas propose three criteria for effective labels and report experimental findings based on 300 users.

PART III (*Automatic Web Genre Identification*) presents the state of the art in automatic genre identification based on the traditional vector space approach. This part includes five chapters, each one showing how automatic genre identification is needed in a wide range of disciplines, and can be achieved with a wide range of features.

In computational linguistics, Santini highlights the need for evaluating the generality and scalability of genre models. For this reason, she suggests using cross-testing techniques, while optimistically waiting for the construction of a genre reference corpus.

Kim and Ross present powerful features that perform well with a large number of genres, which have been selected for digital library applications.

In corpus linguistics, Sharoff is looking for a genre palette and genre model that can permit comparisons between traditional corpora and web corpora. He proposes seven functional genre categories that could be applied to virtually any text found on the Web.

Stein and co-workers present implementation aspects for a genre-enabled web search. They focus on the generalisation capability of web genre retrieval models, for which they propose new evaluation measures and a quantitative analysis.

Braslavski studies the effects of aggregating genre-related and text relevance rankings. His results show moderate positive effects, and encourage further research in this direction.

PART IV (*Structure-oriented Models of Web Genres*) focuses on genres at the website or network level, where structural information play a primary role.

Lindemann and Littig propose a vector-space approach for the automatic identification of super-genres at website level with excellent results.

Dehmer and Emmert-Streib discuss a graph-based perspective for automatically analysing web genre data by mining graph patterns representing web-based hypertext structures. The contribution emphasises how an approach entirely different from the vector space model can be effective.

Björneborn outlines an exploratory empirical investigation of genre connectivity in an academic web space, i.e., how web page genres are connected by links. The pages are categorised into nine institutional and eight personal genre classes. The author builds a genre network graph to discuss changes in page genres and page topics along link paths.

PART V (*Case Studies of Web Genres*) focuses on the empirical observation of emerging web genres.

Paolillo and co-workers apply the social network approach to detect genre emergence in the amateur Flash community by observing social interaction. Their results indicate that participants' social network positions are strongly associated with the genres of Flash they produce, and this contributes to the establishment of genre norms.

Grieve and co-workers apply Biber's multi-dimensional analysis to investigate functional linguistic variation in Internet blogs, with the goal of identifying text types that are distinguished linguistically. Two main sub-types of blogs are identified: personal blogs and thematic blogs.

Bruce first reviews approaches to the notion of genre as a method of categorisation of written texts, leading to the presentation of a rationale for the dual approach of social genre and cognitive genre. This approach is used to analyse ten sample texts of the participatory journalism genre. The author concludes by saying that an adequate operationalisation of a genre as a category of written texts, including a web genre, should be able to account for the socially-constructed cognitive, organisational, and linguistic elements of genre knowledge.

References

[1] Einat Amitay, David Carmel, Adam Darlow, Ronny Lempel, and Aya Soffer. The connectivity sonar: detecting site functionality by structural patterns. In *Proc. of the 14th ACM conference on Hypertext and Hypermedia*, pages 38–47, 2003.

[2] Jack Andersen. The concept of genre in information studies. *Annual Review of Information Science & Technology, Volume 42: 2008*, page 339, 2007.

[3] Jack Andersen. Bringing genre into focus: Lis and genre between people, texts, activity and situation. *Bulletin of the American Society for Information Science and Technology*, 34(5), 2008.

[4] Inger Askehave and Anne E. Nielsen. Digital genres: a challenge to traditional genre theory. *Information Technology & People*, 18(2):120–141, 2005.

[5] Guy Aston and Lou Burnard. *The BNC Handbook: Exploring the British National Corpus with SARA*. Edinburgh University Press, Edinburgh, 1998.

[6] David T. Barnard, Lou Burnard, Steven J. DeRose, David G. Durand, and C. Michael Sperberg-McQueen. Lessons for the World Wide Web from the text encoding initiative. In *Proc. of the 4th International World Wide Web Conference "The Web Revolution"*, Boston, Massachusetts, 1995.

[7] Marco Baroni and Adam Kilgarriff. Large linguistically-processed Web corpora for multiple languages. In *Companion Volume to Proc. of the European Association of Computational Linguistics*, pages 87–90, Trento, 2006.

[8] Marco Baroni, Francis Chantree, Adam Kilgarriff, and Serge Sharoff. Cleaneval: a competition for cleaning web pages. In *Proc. of the Sixth Language Resources and Evaluation Conference, LREC 2008*, Marrakech, 2008.

[9] John A. Bateman. *Multimodality and genre: A foundation for the systematic analysis of multimodal documents*. Palgrave Macmillan, 2008.

[10] John A. Bateman, Thomas Kamps, Jörg Kleinz, and Klaus Reichenberger. Towards constructive text, diagram, and layout generation for information presentation. *Computational Linguistics*, 27(3):409–449, 2001.

[11] Douglas Biber. *Variation across speech and writing*. Cambridge University Press, 1988.

[12] Douglas Biber. A typology of English texts. *Linguistics*, 27(3):43–58, 1989.

[13] Douglas Biber. *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge University Press, Cambridge, 1995.

[14] Douglas Biber, Ulla Connor, and Thomas A. Upton. *Discourse on the move: using corpus analysis to describe discourse structure*. Benjamins, 2007.

[15] Lennart Björneborn. *Small-World Link Structures across an Academic Web Space: A Library and Information Science Approach*. PhD thesis, Royal School of Library and Information Science, Department of Information Studies, Denmark, 2004.

[16] Lennart Björneborn. Genre connectivity and genre drift in a web of genres. In Mehler et al. [68].

[17] Pavel Braslavski. Marrying relevance and genre rankings: an exploratory study. In Mehler et al. [68].

[18] Ian Bruce. *Academic writing and genre: a systematic analysis*. Continuum, 2008.

[19] Ian Bruce. Evolving genres in online domains: The hybrid genre of the participatory news article. In Mehler et al. [68].

[20] Soumen Chakrabarti. Integrating the document object model with hyperlinks for enhanced topic distillation and information extraction. In *Proc. of the 10th International World Wide Web Conference, Hong Kong, May 1-5*, pages 211–220, 2001.

[21] Soumen Chakrabarti, Martin van den Berg, and Byron Dom. Focused crawling: a new approach to topic-specific web resource discovery. In *Proc. 8^{th} International World Wide Web Conference*, Toronto, 1999.

[22] Soumen Chakrabarti, Mukul Joshi, Kunal Punera, and David M. Pennock. The structure of broad topics on the web. In *Proc. of the 11th Internat. World Wide Web Conference*, pages 251–262. ACM Press, 2002.

[23] David A. Cohn and Thomas Hofmann. The missing link – a probabilistic model of document content and hypertext connectivity. In Todd K. Leen, Thomas G. Dietterich, and Volker Tresp, editors, *Advances in Neural Information Processing Systems 13, Papers from Neural Information Processing Systems (NIPS)*, pages 430–436. MIT Press, 2000.

[24] Anne Condamines. Taking genre into account when analysing conceptual relation patterns. *Corpora*, 3(2):115–140, 2008.

[25] Mark Craven, Dan DiPasquo, Dayne Freitag, Andrew K. McCallum, Tom M. Mitchell, Kamal Nigam, and Seán Slattery. Learning to construct knowledge bases from the World Wide Web. *Artificial Intelligence*, 118(1-2):69–113, 2000.

[26] Matthias Dehmer and Frank Emmert-Streib. Mining graph patterns in web-based systems: A conceptual view. In Mehler et al. [68].

[27] Ludovic Denoyer and Patrick Gallinari. Un modèle de mixture de modèles génératifs pour les documents structurés multimédias. *Document numérique*, 8(3), 2004.

[28] Michelangelo Diligenti, Marco Gori, Marco Maggini, and Franco Scarselli. Classification of HTML documents by hidden tree-markov models. In *Proc. of the International Conference on Document Analysis and Recognition (ICDAR), Seattle, WA*, pages 849–853, 2001.

[29] Andrew Dillon. Bringing genre into focus: Why information has shape. *Bulletin of the American Society for Information Science and Technology*, 34(5), 2008.

[30] Debora Donato, Luigi Laura, Stefano Leonardi, and Stefano Millozzi. The web as a graph: How far we are. *ACM Trans. Inter. Tech.*, 7(1), 2007.

[31] Nadav Eiron and Kevin S. McCurley. Untangling compound documents on the web. In *Proc. of the 14th ACM conference on Hypertext and Hypermedia, Nottingham, UK*, pages 85–94, 2003.

[32] Martin Ester, Hans-Peter Kriegel, and Matthias Schubert. Web site mining: a new way to spot competitors, customers and suppliers in the world wide web. In *KDD '02: Proc. of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 249–258, New York, NY, USA, 2002. ACM.

[33] Adriano Ferraresi, Eros Zanchetta, Silvia Bernardini, and Marco Baroni. 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, 2008.

[34] William H. Fletcher. Making the web more useful as a source for linguistic corpora. In U. Connor and T. Upton, editors, *Corpus Linguistics in North America 2002: Selections from the Fourth North American Symposium of the American Association for Applied Corpus Linguistics*, 2004.

[35] Paolo Frasconi, Giovanni Soda, and Alessandro Vullo. Hidden markov models for text categorization in multi-page documents. *Journal of Intelligent Information Systems*, 18(2-3):195–217, 2002.

[36] Luanne Freund. *Exploiting task-document relationships to support information retrieval in the workplace*. PhD thesis, University of Toronto, 2008.

[37] Luanne Freund and Christina Nilsen. Assessing a genre-based approach to online government information. *Proceedings of the 36th annual conference of the Canadian Association for Information Science (CAIS)*, 2008.

[38] Jack Grieve, Douglas Biber, Eric Friginal, and Tatiana Nekrasova. Variation among blogs: A multi-dimensional analysis. In Mehler et al. [68].

[39] Mikael Gunnarsson. *Classification along Genre Dimensions*. Phd, Inst. f. Biblioteks- och Informationsvetenskap. Göteborgs Universitet, 2010.

[40] Suhit Gupta, Hila Becker, Gail Kaiser, and Salvatore Stolfo. Verifying genre-based clustering approach to content extraction. In *Proceedings of the 15th international conference on World Wide Web*, pages 875–876. ACM, 2006.

[41] Bin He, Mitesh Patel, Zhen Zhang, and Kevin Chen-Chuan Chang. Accessing the deep web: A survey. *Communications of the ACM*, 50(2): 94–101, 2007.

[42] Susan C. Herring, Inna Kouper, John C. Paolillo, Lois Ann Scheidt, Michael Tyworth, Peter Welsch, Elijah Wright, and Ning Yu. Conversations in the blogosphere: An analysis "from the bottom up". In *Proc. of the 38th Annual Hawaii International Conference on System Sciences (HICSS'05)*, 2005.

[43] Theresa Heyd. *Email hoaxes: form, function, genre ecology*. Benjamins, 2008.

[44] Nancy Ide, Randi Reppen, and Keith Suderman. The American National Corpus: More than the Web can provide. In *Proc. of the Third Language Resources and Evaluation Conference*, pages 839–844, Las Palmas, 2002.

[45] Thorsten Joachims, Nello Cristianini, and John Shawe-Taylor. Composite kernels for hypertext categorisation. In *Proc. of the 11th International Conference on Machine Learning*, pages 250–257. Morgan Kaufmann, 2001.

[46] Ioannis Kanaris and Efstathios Stamatatos. Webpage genre identification using variable-length character n-grams. In *Proc. of the 19th IEEE Int. Conf. on Tools with Artificial Intelligence (ICTAI'07)*, Washington, DC, USA, 2007. IEEE Computer Society.

[47] Jussi Karlgren and Douglass Cutting. Recognizing text genres with simple metrics using discriminant analysis. In *Proceedings of the 15th conference on Computational linguistics-Volume 2*, pages 1071–1075, 1994.

[48] Brett Kessler, Geoffrey Nunberg, and Hinrich Schütze. Automatic detection of text genre. *Proc. of ACL*, 1997.

[49] Yunhyong Kim and Seamus Ross. Formulating representative features with respect to genre classification. In Mehler et al. [68].

[50] Hans-Peter Kriegel and Matthias Schubert. Classification of websites as sets of feature vectors. In M. H. Hamza, editor, *Databases and Applications*, pages 127–132. IASTED/ACTA Press, 2004.

[51] Ravi Kumar, Jasmine Novak, Prabhakar Raghavan, and Andrew Tomkins. Structure and evolution of blogspace. *Communications of the ACM*, 47(12):35–39, 2004.

[52] Henry Kučera and W. Nelson Francis. *Computational analysis of present-day American English*. Brown University Press, Providence, 1967.

[53] David Lee. 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, 2001.

[54] Wen-Syan Li, Okan Kolak, Quoc Vu, and Hajime Takano. Defining logical domains in a web site. In *Proc. of the 11th ACM on Hypertext and Hypermedia*, pages 123–132, 2000.

[55] Wen-Syan Li, K. Selçuk Candan, Quoc Vu, and Divyakant Agrawal. Query relaxation by structure and semantics for retrieval of logical web documents. *IEEE Transactions on Knowledge and Data Engineering*, 14 (4):768–791, 2002.

[56] Chul Su Lim, Kong Joo Lee, and Gil Chang Kim. Multiple sets of features for automatic genre classification of web documents. *Information Processing & Management*, 41(5):1263–1276, 2005.

[57] Christoph Lindemann and Lars Littig. Classification of web sites at super-genre level. In Mehler et al. [68].

[58] J. Luzon, N. Ruiz, and Villanueva L. *Genre Theory and new literacies. Applications to autonomous language learning*. Springer, In preparation.

[59] Elizabeth Marshman, Marie-Claude L'Homme, and Victoria Surtees. Portability of cause-effect relation markers across specialised domains and text genres: a comparative evaluation. *Corpora*, 3(2):141–172, 2008.

[60] James R. Martin. Macro-genres: the ecology of the page. *Network*, 21: 29–52, 1994.

[61] James R. Martin and David Rose. *Genre Relations: mapping culture*. Equinox Pub., 2008.

[62] Alexander Mehler. Structural similarities of complex networks: A computational model by example of wiki graphs. *Applied Artificial Intelligence*, 22(7&8):619–683, 2008.

[63] Alexander Mehler. Structure formation in the web. A graph-theoretical model of hypertext types. In Andreas Witt and Dieter Metzing, editors, *Linguistic Modeling of Information and Markup Languages. Contributions to Language Technology*, Text, Speech and Language Technology. Springer, Dordrecht, 2009.

[64] Alexander Mehler. Generalised shortest paths trees: A novel graph class applied to semiotic networks. In Matthias Dehmer and Frank-Emmert Streib, editors, *Analysis of Complex Networks: From Biology to Linguistics*. Wiley-VCH, Weinheim, 2009.

[65] Alexander Mehler. A quantitative graph model of social ontologies by example of Wikipedia. In Matthias Dehmer, Frank Emmert-Streib, and Alexander Mehler, editors, *Towards an Information Theory of*

Complex Networks: Statistical Methods and Applications. Birkhäuser, Boston/Basel, 2009.

[66] Alexander Mehler, Matthias Dehmer, and Rüdiger Gleim. Towards logical hypertext structure: a graph-theoretic perspective. In Th. Böhme and G. Heyer, editors, *Proc. of the 4th International Workshop on Innovative Internet Computing Systems (I2CS '04)*, Lecture Notes in Computer Science 3473, pages 136–150, Berlin/New York, 2006. Springer.

[67] Alexander Mehler, Rüdiger Gleim, and Armin Wegner. Structural uncertainty of hypertext types. An empirical study. In *Proc. of the Workshop "Towards Genre-Enabled Search Engines: The Impact of NLP", September, 30, 2007, in conjunction with RANLP 2007, Borovets, Bulgaria*, pages 13–19, 2007.

[68] Alexander Mehler, Serge Sharoff, and Marina Santini, editors. *Genres on the Web: Computational Models and Empirical Studies*. Springer, Berlin/New York, 2009.

[69] Filippo Menczer. Lexical and semantic clustering by web links. *Journal of the American Society for Information Science and Technology*, 55(14): 1261–1269, 2004.

[70] Michela Montesi and Trilce Navarrete. Classifying web genres in context: A case study documenting the web genres used by a software engineer. *Information Processing and Management*, 2008.

[71] Iadh Ounis, Maarten De Rijke, Craig Macdonald, Gilad Mishne, and Ian Soboroff. Overview of the trec 2006 blog track. In *Proc. of the Text Retrieval Conference (TREC), NIST*, 2006.

[72] Tero Päivärinta, Michael Shepherd, Lars Svensson, and Matti Rossi. A special issue editorial. *Scandinavian Journal of Information Systems*, 20 (1), 2008.

[73] Peter Pirolli, James Pitkow, and Ramana Rao. Silk from a sow's ear: Extracting usable structures from the web. In *Proc. of the ACM SIGCHI Conference on Human Factors in Computing*, pages 118–125. ACM Press, 1996.

[74] Richard Power, Donia Scott, and Nadjet Bouayad-Agha. Document structure. *Computational Linguistics*, 29(2):211–260, 2003.

[75] Tapani Raiko, Kristian Kersting, Juha Karhunen, and Luc De Raedt. Bayesian learning of logical hidden markov models. In *Proc. of the Finnish AI conference (STeP-2002)*, pages 64–71, 2002.

[76] Georg Rehm. 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*, 2002.

[77] Georg Rehm. Hypertext types and markup languages. the relationship between HTML and web genres. *Linguistic Modelling of Information and Markup Languages*, 2008.

[78] Mark A. Rosso. Bringing genre into focus: Stalking the wild web genre (with apologies to euell gibbons). *Bulletin of the American Society for Information Science and Technology*, 34(5), 2008.

[79] Mark A. Rosso and Stephanie W. Haas. Identification of web genres by user warrant. In Mehler et al. [68].

[80] Marina Santini. Characterizing genres of web pages: Genre hybridism and individualization. In *Proc. of the 40th Annual Hawaii International Conference on System Sciences (HICSS'07)*, 2007.

[81] Marina Santini. *Automatic Identification of Genre in Web Pages*. Phd thesis, University of Brighton, Brighton, United Kingdom, 2007.

[82] Marina Santini. Cross-testing a genre classification model for the web. In Mehler et al. [68].

[83] Serge Sharoff. Creating general-purpose corpora using automated search engine queries. In Marco Baroni and Silvia Bernardini, editors, *WaCky! Working papers on the Web as Corpus*. Gedit, Bologna, 2006.

[84] Serge Sharoff. Classifying web corpora into domain and genre using automatic feature identification. In *Proc. of Web as Corpus Workshop*, Louvain-la-Neuve, 2007.

[85] Serge Sharoff. In the garden and in the jungle. Comparing genres in the bnc and internet. In Mehler et al. [68].

[86] John Sinclair, editor. *Looking up: an account of the COBUILD Project in lexical computing*. Collins, London and Glasgow, 1987.

[87] John Sinclair. Corpora for lexicography. In P. van Sterkenberg, editor, *A Practical Guide to Lexicography*, pages 167–178. Benjamins, Amsterdam, 2003.

[88] Benno Stein, Sven Meyer zu Eissen, and Nedim Lipka. Web genre analysis: Use cases, retrieval models, and implementation issues. In Mehler et al. [68].

[89] Jade G. Stewart. *Genre Oriented Summarization*. PhD thesis, Carnegie Mellon University, 2008.

[90] Aixin Sun and Ee-Peng Lim. Web unit mining: finding and classifying subgraphs of web pages. In *CIKM '03: Proc. of the twelfth international conference on Information and knowledge management*, pages 108–115, New York, NY, USA, 2003. ACM.

[91] John M. Swales. *Genre analysis: English in academic and research settings*. Cambridge University Press, 1990.

[92] Keishi Tajima and Katsumi Tanaka. New techniques for the discovery of logical documents in web. In *Internat. Symposium on Database Applications in Non-Traditional Environments*, pages 125–132. IEEE, 1999.

[93] Keishi Tajima, Yoshiaki Mizuuchi, Masatsugu Kitagawa, and Katsumi Tanaka. Cut as a querying unit for WWW, netnews, e-mail. In *Proc. of the 9th ACM Conference on Hypertext and Hypermedia*, pages 235–244. ACM, 1998.

[94] Mike Thelwall, Liwen Vaughan, and Lennart Björneborn. Webometrics. *Annual Review of Information Science Technology*, 6(8), 2006.

[95] Yong Hong Tian, Tie Jun Huang, Wen Gao, Jun Cheng, and Ping Bo Kang. Two-phase web site classification based on hidden markov tree models. In *WI '03: Proc. of the 2003 IEEE/WIC International Conference on Web Intelligence*, page 227, Washington, DC, USA, 2003. IEEE Computer Society.

[96] Ulli Waltinger, Alexander Mehler, and Armin Wegner. A two-level approach to web genre classification. In *Proc. of the 5th International Conference on Web Information Systems and Technologies (WEBIST '09), March 23-26, 2007, Lisboa*, 2009.

[97] Guillaume Wisniewski, Francis Maes, Ludovic Denoyer, and Patrick Gallinari. Modèle probabiliste pour l'extraction de structures dans les documents web. *Document numérique*, 10(1), 2007.

[98] Ruth Wodak. Introduction: Discourse studies - important concepts and terms. *Qualitative discourse analysis in the Social Sciences*, 2008.

[99] Simeon J. Yates and Tamara R. Sumner. Digital genres and the new burden of fixity. In *System Sciences, 1997, Proceedings of the Thirtieth Hawaii International Conference on*, volume 6, 1997.