Genres on the Web
Text, Speech and Language Technology

VOLUME 42

Series Editors

Nancy Ide, Vassar College, New York
Jean Véronis, Université de Provence and CNRS, France

Editorial Board

Harald Baayen, Max Planck Institute for Psycholinguistics, The Netherlands
Kenneth W. Church, Microsoft Research Labs, Redmond WA, USA
Judith Klavans, Columbia University, New York, USA
David T. Barnard, University of Regina, Canada
Dan Tufis, Romanian Academy of Sciences, Romania
Joaquim Llisterri, Universitat Autonoma de Barcelona, Spain
Stig Johansson, University of Oslo, Norway
Joseph Mariani, LIMSI-CNRS, France

For further volumes:
http://www.springer.com/series/6636
As a reader, I’m looking for two things from a new book on genre. First, does it offer some new tools for analysing genres; and second, does it explore genres that haven’t been much studied before? Genres on the Web delivers brilliantly on both accounts, introducing as it does a host of computational perspectives on genre classification and focussing as it does on a range of newly emerging electronic genres. Lacking expertise in the computational modelling thematised throughout the book I can’t do much more here than express my fascination with the questions tackled and methods deployed. Having expertise in functional linguistics and its deployment in genre-based literacy programs I can perhaps offer a few observations that might help push this and comparable endeavours along.

First some comments as a functional linguist. Characterising almost all the papers is a two-level approach nicely summarised by Stein et al. in their Table 8.1. On the one hand we have a web genre palette, with many alternative classifications of genres; on the other hand we have document representation, with the many alternative sets of features used to explore web data in relation to genre. The most striking thing about this perspective to me is its relatively flat approach as far as social context and its realisation in language and attendant modalities of communication is concerned.

In systemic functional linguistics for example, it is standard practice to explore variation across texts from the perspectives of field, tenor and mode as well as genre. Field is concerned with institutional practice – domestic activity, sport and recreation, administration and technology, science, social science and humanities and so on. Tenor is concerned with social relations negotiated – in relation to power (equal/unequal) and solidarity (intimate, collegial, professional etc.). Mode is concerned with the affordances of the channel of communication – how does the technology affect interactivity (both type and immediacy), degree of abstraction (e.g. texts accompanying physical behaviour, recounting it, reflecting on it, theorising it) and intermodality (the contribution of language, image, sound, gesture etc. to the text at hand). In my own work genre is then deployed to describe how a culture combines field, tenor and mode variables into recurrent configurations of meaning and phases these into the unfolding stages typifying that social process.

When I referred to a flat model of social context above what I meant was that in this book these four contextual variables tend to be conflated into a single taxonomy of text types, without there being any apparent theoretically informed set of
principles for the flattening. It may well be of course that for one reason or another we do want a simple model of social context and may wish to foreground one field or mode or tenor variable over another. But it might prove more useful to begin with a richer theory of context than we need for any one task, and flatten it in principle, than to try and build a parsimonious model from the start, and complicate it over time.

Turning to document representation, once again from the perspective of systemic functional linguistics, it is standard practice to explore representation in language (and other modalities of communication) from the perspective of various hierarchies and complementarities. The chief hierarchies used are rank (how large are the units considered – e.g. word, phrase, clause, phase, stage, text) and strata (which level of abstraction from materiality is being considered – phonology/graphology, lexicogrammar or discourse semantics). The chief complementarity used is metafunction (are we considering the ideational meanings used to naturalise a picture of reality, the interpersonal meanings used to negotiate social relationships or the textual meanings used to weave these together as waves of information in interpretable discourse).

The meanings dispersed across these ranks, strata and metafunctions are regularly collapsed into a list of descriptive features in this volume, when for different purposes one might want to be selective or value some features over others. Exacerbating this is an apparent need to foreground relatively low-level formal features which are easily computable, since manual analysis is too slow and costly, and in any case so much of the research here is focussed on the automatic retrieval of genres. Beyond this, as Kim and Ross point out, texts are regularly treated as bags of features, as if the timing of their realisation plays no significant part in the recognition of a genre. What saddens me here is the gulf between computational and linguistically informed modelling of genres, for which I know my colleagues in linguistics are responsible – since for the most part they work on form not meaning, and focus on the form of clauses and syllables, not discourse (they still think a language is a set of sentences rather than a communication system instantiated through an indefinitely large lattice of texts).

Next some comments as a functional linguist working in language and education programs over three decades. From the start we of course faced the problem of classifying texts – in our case the genres that students needed to read and write in primary, secondary and tertiary sectors of education, and their relation to workplace discourse and professional development therein. One thing we learned from this work was to be wary of the folk-classifications of genres used by educators. Our primary school teachers for example called everything their students wrote a story, when in fact, from a linguistic perspective, the students engaged in a range of genres. Complicating this was their tendency to evaluate everything the students wrote as a story, in spite of suggesting to students that they choose their own topics or even that they write in any form they choose. As an issue of social justice, we felt we had to replace the folk-categorisation with a linguistically informed one, and take the further step of insisting that this uncommon sense classification be shared between teachers and students. The moral of this experience I feel is that we need to treat
“folksonomies” with great caution when classifying genres, and not expect users to be able to easily bring to consciousness or even demonstrate in practice a genre classification that will best suit the purposes of our own research.

Throughout this literacy focussed action research we have lacked the funding and computational tools to undertake the systematic quantitative analysis thematised in this volume. Instead we had to rely on manual analysis of texts our teacher linguists selected as representative (depending as they did on their own experience, advice from teachers, assessment processes and textbook exemplars). This meant we could build up a picture of genres based on thick descriptions of all the levels of analysis I worried about being flattened above; the great weakness of this approach of course is replicability – were our few texts in fact representative and would quantitative analysis support our findings over time? In practice, the only confirmation we received that we were on the right track lay in the literacy progress of our students, since we were interested in genre because we wanted to redistribute the meaning potential of our culture more evenly than schools have been able to do in the past.

At this point I suspect that most of the authors in this volume would throw up their hands in despair of finding anything useful in our work. So let me just end on a note of caution. What if genres cannot be robustly characterised on the basis of just a few easily computable formal features? What if a flat approach to contextual variables and representational features simplifies research to the point where it is hard to see how the texts considered could have evolved as realisations of the genres members of our culture use to live? Would we be wise to complement flat computationally based quantitative analysis with thick manual qualitative description and see where the two trajectories lead us? And do we need to balance commercially driven research with ideologically committed initiatives (who for example will benefit from the genre informed search engines inspiring so many of the papers herein)?

I’ll stop here, concerned that this preface is turning into a post-script, or even a chapter in a book where prefacing is where I barely belong! My thanks to the editors for opening up this work, which will prove indispensable for readers with many converging concerns. I’ll do what I can to point my students and colleagues in the direction of the transdisciplinary dialogue which I’m sure will be inspired by the genre analysts dialoguing here.

Sydney, Australia
March 2009

James R. Martin
Personal Note

Here let us breathe and haply institute
A course of learning and ingenious studies.
Shakespeare, The taming of the shrew, Act I, scene I

To all of you who have been involved in this book I want to say: Thank you! This book is very much the result of your collective efforts. It would not have come about without your commitment and interest in the concept of genre, this untamed shrew.

My first mention goes to the authors who readily accepted to contribute to this volume. Many thanks for your chapters, dear Authors, that show the state of the art of empirical and computational genre research.

I am also most grateful to our reviewers whose comments were most valuable. Many thanks for your detailed feedback, dear Reviewers, that has improved the content, presentation and style of our chapters.

Thank you to everybody for sharing your knowledge and dedication to make this volume possible.

Have we started taming the shrew? I am sure we have.

Marina Santini
Book Coordinator
Contents

Part I  Introduction

1  Riding the Rough Waves of Genre on the Web  ......................... 3
    Marina Santini, Alexander Mehler, and Serge Sharoff

Part II  Identifying the Sources of Web Genres

2  Conventions and Mutual Expectations  .......................... 33
    Jussi Karlsgren

3  Identification of Web Genres by User Warrant  ...................... 47
    Mark A. Rosso and Stephanie W. Haas

4  Problems in the Use-Centered Development of a Taxonomy
    of Web Genres  ................................................ 69
    Kevin Crowston, Barbara Kwaśnik, and Joseph Rubleske

Part III  Automatic Web Genre Identification

5  Cross-Testing a Genre Classification Model for the Web .......... 87
    Marina Santini

6  Formulating Representative Features with Respect to Genre
    Classification  .................................................. 129
    Yunhyong Kim and Seamus Ross

7  In the Garden and in the Jungle  ................................. 149
    Serge Sharoff
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td><strong>Web Genre Analysis: Use Cases, Retrieval Models, and Implementation Issues</strong></td>
<td>167</td>
</tr>
<tr>
<td></td>
<td>Benno Stein, Sven Meyer zu Eissen, and Nedim Lipka</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td><strong>Marrying Relevance and Genre Rankings: An Exploratory Study</strong></td>
<td>191</td>
</tr>
<tr>
<td></td>
<td>Pavel Braslavski</td>
<td></td>
</tr>
</tbody>
</table>

**Part IV Structure-Oriented Models of Web Genres**

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td><strong>Classification of Web Sites at Super-Genre Level</strong></td>
<td>211</td>
</tr>
<tr>
<td></td>
<td>Christoph Lindemann and Lars Littig</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td><strong>Mining Graph Patterns in Web-Based Systems: A Conceptual View</strong></td>
<td>237</td>
</tr>
<tr>
<td></td>
<td>Matthias Dehmer and Frank Emmert-Streib</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td><strong>Genre Connectivity and Genre Drift in a Web of Genres</strong></td>
<td>255</td>
</tr>
<tr>
<td></td>
<td>Lennart Björneborn</td>
<td></td>
</tr>
</tbody>
</table>

**Part V Case Studies of Web Genres**

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td><strong>Genre Emergence in Amateur Flash</strong></td>
<td>277</td>
</tr>
<tr>
<td></td>
<td>John C. Paolillo, Jonathan Warren, and Breanne Kunz</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td><strong>Variation Among Blogs: A Multi-Dimensional Analysis</strong></td>
<td>303</td>
</tr>
<tr>
<td></td>
<td>Jack Grieve, Douglas Biber, Eric Friginal, and Tatiana Nekrasova</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td><strong>Evolving Genres in Online Domains: The Hybrid Genre of the Participatory News Article</strong></td>
<td>323</td>
</tr>
<tr>
<td></td>
<td>Ian Bruce</td>
<td></td>
</tr>
</tbody>
</table>

**Part VI Prospect**

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td><strong>Any Land in Sight?</strong></td>
<td>351</td>
</tr>
<tr>
<td></td>
<td>Marina Santini, Serge Sharoff, and Alexander Mehler</td>
<td></td>
</tr>
</tbody>
</table>

**Index**                                                                 | 355  |
Contributors

Douglas Biber  English Department, Northern Arizona University, Flagstaff, AZ, USA, douglas.biber@nau.edu

Lennart Björneborn  Royal School of Library and Information Science, Copenhagen, Denmark, lb@iva.dk

Pavel Braslavski  Institute of Engineering Science RAS, 620219 Ekaterinburg, Russia, pb@imach.uran.ru; pb@yandex-team.ru

Ian Bruce  University of Waikato, Hamilton, New Zealand, ibruce@waikato.ac.nz

Kevin Crowston  School of Information Studies, Syracuse University, Syracuse, NY, USA, crowston@syr.edu

Matthias Dehmer  Institute of Discrete Mathematics and Geometry, Vienna University of Technology, Vienna, Austria; Institute for Bioinformatics and Translational Research, Hall in Tyrol, Austria, matthias.dehmer@univie.ac.at; mdehmer@geometrie.tuwien.ac.at; Matthias.Dehmer@umit.at

Frank Emmert-Streib  Computational Biology and Machine Learning, Center for Cancer Research and Cell Biology, School of Medicine, Dentistry and Biomedical Sciences, Queen’s University Belfast, Belfast, UK, f@bio-complexity.com

Eric Friginal  Department of Applied Linguistics and English as a Second Language, Georgia State University, Atlanta, GA, USA, efriginal@gsu.edu

Jack Grieve  QLVL Research Unit, University of Leuven, Leuven, Belgium, Jack.Grieve@arts.kuleuven.be

Stephanie W. Haas  School of Information & Library Science, University of North Carolina, Chapel Hill, NC 27599-3360, USA, shaas@email.unc.edu

Jussi Karlgren  Swedish Institute of Computer Science (SICS), Stockholm, Sweden, jussi@sics.se

Yunhyong Kim  Humanities Advanced Technology and Information Institute (HATII), University of Glasgow, Glasgow, UK; School of Computing, Robert Gordon University, Aberdeen, UK, ykim1@rgu.ac.uk
Breanne Kunz  School of Library and Information Science and School of Informatics, Indiana University, Bloomington, IN 47408, USA, b kunz@indiana.edu

Barbara Kwasnik  School of Information Studies, Syracuse University, Syracuse, NY, USA, bkwasnik@syr.edu

Christoph Lindemann  Department of Computer Science, University of Leipzig, Leipzig, Germany, cl@rvs.informatik.uni-leipzig.de

Nedim Lipka  Faculty of Media/Media Systems, Bauhaus-Universität Weimar, Weimar, Germany, nedim.lipka@uni-weimar.de

Lars Littig  Department of Computer Science, University of Leipzig, Leipzig, Germany, littig@rvs.informatik.uni-leipzig.de

Alexander Mehler  Computer Science and Mathematics, Goethe-Universität Frankfurt am Main, Georg-Voigt-Straße 4, D-60325 Frankfurt am Main, Germany, Mehler@em.uni-frankfurt.de

Sven Meyer zu Eissen  Faculty of Media/Media Systems, Bauhaus-Universität Weimar, Weimar, Germany, sven@meyer-zu-eissen.de; sven.meyer-zu-eissen@medien.uni-weimar.de

Tatiana Nekrasova  English Department, Northern Arizona University, Flagstaff, AZ, USA, Tatiana.Nekrasova@nau.edu

John C. Paolillo  School of Library and Information Science and School of Informatics, Indiana University, Bloomington, IN 47408, USA, paolillo@indiana.edu

Seamus Ross  iSchool, University of Toronto, Toronto, CA, seamus.ross@utoronto.ca

Mark A. Rosso  School of Business, North Carolina Central University, Durham, NC 27707, USA, mrosso@nccu.edu

Joseph Rubleske  School of Information Studies, Syracuse University, Syracuse, NY, USA, jrublesk@gmail.com

Marina Santini  KYH, Stockholm, Sweden, marinasantini.ms@gmail.com

Serge Sharoff  Centre for Translation Studies, University of Leeds, LS2 9JT Leeds, UK, s.sharoff@leeds.ac.uk

Benno Stein  Faculty of Media/Media Systems, Bauhaus-Universität Weimar, Weimar, Germany, benno.stein@uni-weimar.de

Jonathan Warren  School of Library and Information Science and School of Informatics, Indiana University, Bloomington, IN 47408, USA, jowarren@indiana.edu
Part I
Introduction
Chapter 1
Riding the Rough Waves of Genre on the Web
Concepts and Research Questions

Marina Santini, Alexander Mehler, and Serge Sharoff

1.1 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.¹

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

---

¹ 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.
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.

1.1.1 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,\(^2\) in the identification of context of interaction and search,\(^3\) in the identification of the “sentiment” contained in a text,\(^4\) and in other aspects affecting the reliability, trust, reputation\(^5\) 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 Karlsgren and Cutting [47] and Kessler et al. [48]. Rosso [76] 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 (see Chapter 8 by Stein et al., 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. [96]), 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. [14, 24, 58]), automatic summarisation [87], information extraction [40], creation of language corpora [82], e-government (e.g. [37]), information science (e.g. [39] or [68]), information systems [70] and many other activities.

The genres used by Karlsgren 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

\(^2\) For instance, see “PAN’09: 3rd Int. PAN Workshop – 1st Competition on Plagiarism Detection”.
\(^3\) For instance, see “ECIR 2009 Workshop on Contextual Information Access, Seeking and Retrieval Evaluation”.
\(^4\) For instance, see “CyberEmotions” http://www.cyberemotions.eu/
\(^5\) For instance, see “WI/IAT’09 Workshop on Web Personalization, Reputation and Recommender Systems”.
in genre research (see Chapter 3 by Rosso and Haas; Chapter 4 by Crowston et al., this volume). But, after all, the naming difficulty is very much connected with the arduousness of defining genre and characterising genre classes.

1.2 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. cyber-genre).

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)\(^6\) – 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 \(<\text{content, form, functionality}>\), taken as an extension of the original tuple \(<\text{content, form}>\) is misleading (cf. also Karlgren, Chapter 2 in this volume). Also the dual model proposed by Askehave and Nielsen [4], which extends the notion of genre originally developed by Swales [89], 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

\(^6\) The contraposition between these two schools from the perspective of teaching is also well described in Bruce [18], Chapter 2.
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 genres. 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 in Chapter 7 by Sharoff, 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 Chapter 3 by Rosso and Haas; Chapter 4 by Crowston et al., 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 [60] 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 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?

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

Päivärinta et al. [70] 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 to 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? (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, 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. Actually, 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.

### 1.3 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 1.3.1), the construction and use of corpora collected from the web (Section 1.3.2) and the design of computational models (Section 1.3.3).

#### 1.3.1 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 [97] and Rehm [75] 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 single web page? 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, 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 1.1 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, 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 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, although 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

---

7 Global edition: http://www.timesonline.co.uk/tol/global/, or UK edition http://www.timesonline.co.uk/tol/news/
8 As noted by Bateman [9] functionality belongs to both paper and web documents.
to decide which document structures of the web are attributed to web genres: e.g., self-contained pages [78] or their constituents [74, 75, 88, 94], websites [57, 65] 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, 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, this volume). Web genres are manifested by pages [78, 79] 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 [67] 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. [62, 75]):

- **On the micro level** we analyse page-level [77] units and their constituents [88] 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 [69] as instances of this genuine web genre or on networks of blogs which are interlinked by trackbacks and related means [42, 52]. 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 [52]. 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) [61].

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 [80]). 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 [59]. 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. [92] 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 [72] 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 [66]. Manifestations of webgenres are analysed, for example, as compound documents [31], as logical domains [54], as logical documents [55, 91] or as multipage segments [25].9 Whatever is seen to be the exact unit of manifestation

9 See also Tajima et al. [90], Cohn and Hofmann [23] and Chakrabarti et al. [22] for topic-related approaches in this line of research.
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. It might be the case 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. Chapter 4 by Crowston et al., 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 to social, cultural and technological contexts (e.g. see Chapter 13 by Paolillo et al., 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 Fig. 1.1). 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 [64]. 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” up to very recently. 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.
1.3.2 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” [85]. 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 [51]:

- 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 2,000 words (this means that texts themselves had to be longer than 2,000 words);

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, Skills and 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 [84], 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.10

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

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 [81]. 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 4,055 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, 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 web corpora. 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.\footnote{After collecting texts, developers of traditional corpora often introduce their own set of annotation layers, such as POS tagging, semantic or metatextual markup, but such layers are not taken from original texts in the form they have been published.}

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...