

# Metadata of the chapter that will be visualized online

---

ChapterTitle	In the Garden and in the Jungle	
Chapter Sub-Title	Comparing Genres in the BNC and Internet	
Chapter CopyRight - Year	Springer Science+Business Media B.V. 2010 (This will be the copyright line in the final PDF)	
Book Name	Genres on the Web	
Corresponding Author	Family Name	<b>Sharoff</b>
	Particle	
	Given Name	<b>Serge</b>
	Suffix	
	Division	Centre for Translation Studies
	Organization	University of Leeds
	Address	Leeds, UK
	Email	s.sharoff@leeds.ac.uk

---

Abstract	The jungle metaphor is quite common in genre studies. The subtitle of David Lee's seminal paper on genre classification is "navigating a path through the BNC jungle" [16]. According to Adam Kilgarriff, the BNC is a jungle only when compared to smaller Brown-type corpora, while it looks more like an English garden when compared to the Web [15].
----------	---

---

# Chapter 7

## In the Garden and in the Jungle

### Comparing Genres in the BNC and Internet

Serge Sharoff

#### 7.1 Introduction

The jungle metaphor is quite common in genre studies. The subtitle of David Lee's seminal paper on genre classification is "navigating a path through the BNC jungle" [16]. According to Adam Kilgarriff, the BNC is a jungle only when compared to smaller Brown-type corpora, while it looks more like an English garden when compared to the Web [15]. Intuitively this claim is plausible: if we consider the whole Web as a corpus, it probably contains a much greater variety of text types and genres than the 4,055 texts in the BNC classified into 70 genres. However, we still need to study this jungle.

Nowadays it is relatively easy to collect a large corpus from the Web, either using search engines [24] or web crawlers [13, 3], so it is easy to surpass the BNC in size. However, we know little about the domains and genres of texts in corpora collected in this way. Even if we collect domain-specific corpora [2] and can be sure that all texts in our corpus are about, e.g., epilepsy, we still do not know the amount of research papers, newspaper articles, webpages advising parents, tutorials for medical staff, etc, in it.

Traditional corpora have been annotated manually, which did not create a significant overhead: such corpora have been also compiled manually, so it was possible to annotate each text according to a reasonable number of parameters. Even then there can be problems with manual classification. Spoken texts in the BNC are not classified into their domains at all, even though many of them are devoted to a well-defined topic, like computing, medicine or politics. Similarly, a single large text taken from a newspaper and classified as "world affairs" in the BNC can contain home and foreign news, commentaries, gossips, etc. Many genres also remain underdescribed. Even though there are textbooks in the BNC (for instance, texts

---

S. Sharoff (✉)

Centre for Translation Studies, University of Leeds, Leeds, UK

e-mail: s.sharoff@leeds.ac.uk

46 EVW or GVS<sup>1</sup>), their presence is not registered in the classification scheme: they are  
47 classified as written academic texts (according to David Lee's genre classification)  
48 or as books for professional readers in the respective domains of natural sciences and  
49 arts (according to the original BNC database), but nothing in the scheme indicates  
50 that they are teaching materials. However, such complaints are only minor quibbles  
51 if we compare this situation to the sheer lack of information about even very basic  
52 characteristics of Web corpora, such as I-EN [24], SPIRIT [13] or deWaC [3].

53 The task of classifying Web corpora and comparing their composition to tradi-  
54 tional corpora is difficult for several reasons. First, no established classification of  
55 genres exists, even for traditional written texts. Practically every study uses its own  
56 list of genres, e.g., compare the 15 classes in the Brown Corpus to the 70 genres  
57 in David Lee's classification of the BNC to the 120 genre labels in the Russian  
58 National Corpus (RNC). Second, the relationship between traditional genres and  
59 genres existing on the Web is not clear. Some web genres can be compared to tradi-  
60 tional printed media, e.g., on-line newspapers, while others are markedly different  
61 from any known printed counterpart, e.g., chat rooms. Third, given the large num-  
62 ber of pages in Web-derived corpora, e.g., more than 60,000 in I-EN [24], we need  
63 automatic methods that can identify genres reliably and be applicable to an arbitrary  
64 webpage. The fourth problem concerns the very design of the genre inventory. If the  
65 goal is to classify every text existing on the Web, the number of genres is too large  
66 to be listed in a flat list. Only within the genres of academic communication we can  
67 come across research articles (with different genre conventions applicable to the  
68 humanities, engineering or natural sciences), as well as popular articles, reviews,  
69 books, calls for participation, emails, mailing lists and forums, project proposals,  
70 progress reports, minutes of meetings, job descriptions, etc. A recent overview of  
71 traditional genre labels refers to a list of "more than 4,500" categories [21]. The fifth  
72 problem concerns "emerging" genres: new technologies can offer new avenues for  
73 communication, which readily produce new genres, e.g., blogs, personal homepages  
74 or spam. However, we can expect greater stability of underlying communicative  
75 intentions, which are realised in new forms using new technologies. For instance,  
76 if our list of webgenres includes a simple entry for blogs, this category cannot be  
77 compared to anything in the BNC (blogs did not exist at the time of its compilation),  
78 whereas the function of blogs is similar to that of diaries or opinion columns in  
79 newspapers, while it is different from them in the audience size, distribution mode  
80 and authorship.

81 One way of studying genres on the Web is to start with a genre (or a group of gen-  
82 res), such as blogs [17] or conference websites [19]. Then we can analyse linguistic  
83 features specific to this genre and learn how to identify a text as belonging or not  
84 belonging to it. Another way of studying web-genres is aimed at saying "sensible and  
85 useful things about any text" that exists on the Web.<sup>2</sup> Such studies can offer a  
86

---

87  
88 <sup>1</sup> Throughout this chapter I refer to BNC texts using their ids from the BNC Index, which is  
89 available from <http://clix.to/davidlee00>

90 <sup>2</sup> The quote refers to the purposes Michael Halliday intended for his "Introduction to Functional  
Grammar" [11].

## 7 In the Garden and in the Jungle

91 very superficial description for many genres studied in the first approach, but if we  
92 do not use a compact text typology, this study risks ending with an infinite list of  
93 genre types to account for all possible webpages.

94 In this chapter I will follow the latter research path by outlining an approach to  
95 text classification that can be used to describe the majority of texts on the Web using  
96 a small number of categories (less than 10), so that we can broadly assess the com-  
97 position of genres in a Web-derived corpus, compare it against any other collections  
98 of webpages and traditional corpora, as well as against corpora in other languages  
99 (Section 7.2). Then (in Section 7.3) I will present an experiment for detection of  
100 text categories in traditional reference corpora against English and Russian Internet  
101 corpora. Traditional corpora used in this study are the BNC and Russian National  
102 Corpus (RNC), which is comparable to the BNC in its size and composition [23].  
103 Finally, in Section 7.4 I will discuss the similarities and differences between these  
104 Internet corpora and their manually collected counterparts.

105 The study concerns English and Russian corpora collected from the Web using  
106 random queries to search engines [24]. Below these corpora are referred to as “the  
107 Internet corpora” (or I-EN and I-RU more specifically). However, there is nothing  
108 in the methodology specific to this method of corpus collection, so the study should  
109 be applicable to any sufficiently large corpus of webpages (in the discussion below I  
110 refer to such corpora as “Web-derived corpora”). In the last section, I also report on  
111 a small experiment of applying the same methodology to classifying ukWac, another  
112 English corpus collected by crawling websites in the .uk domain [10].

### 7.2 Text Typology for the Web

114  
115  
116  
117 Approaches to classifying texts into genres can be grouped into two main classes.  
118 The first class identifies genres of documents on the basis of what can be called  
119 “look’n’feel” properties, e.g., FAQ, forum or recipe, while the second class detects  
120 broad functional classes, e.g., description or argumentation, cf. the discussion in  
121 [16] or [5].

122 Look’n’feel approaches are based on traditional labels, so they reflect the practice  
123 of their use and it is relatively easy to annotate a significant amount of texts manually  
124 by human annotators without extensive training. For instance, if a page looks like a  
125 blog, applying this label is not difficult for anyone familiar with this genre. If we use  
126 a folksonomy-based genre typology in a search engine, again its users can recognise  
127 labels easily, for instance, to refine the results of their search. At the same time, this  
128 approach assumes an established genre inventory, which does not exist (Problem 1  
129 identified above), it results in proliferation of categories (Problem 4), and it is not  
130 flexible enough to allow comparison of webcorpora to their traditional counterparts  
131 (Problem 5).

132 This is the reason for taking the functional approach to genre classification in  
133 this project. However, even if we narrow our search of a suitable genre classification  
134 scheme down to functional studies, which classify texts from the viewpoint of the  
135 function they fulfill in the society, we still find a large number of options. Marina

136 Santini mentions such classes as Descriptive-narrative, Explicatory-informational,  
137 Argumentative-persuasive and Instructional identified in traditional text typology  
138 studies along with the several variations of this inventory, e.g., separating descriptive  
139 and narrative texts [22, Chapter 2]. Without giving an explicit text typology, James  
140 Martin defines genres as the results of “staged, goal-oriented, purposeful activity in  
141 which speakers engage as members of our culture” [18, p. 25]. In [14], genres are  
142 also defined functionally, but using traditional labels taken from reflective practice,  
143 e.g., “editorial is a shortish prose argument expressing an opinion on some matter of  
144 immediate public concern”. In another study of genre detection [7], the classification  
145 is done into five functional styles: fiction, journalism, official, academic, everyday  
146 language, following a tradition that stems from Jakobson [12].

147 The functional approaches mentioned above are still not precise enough for the  
148 goal of unambiguous classification of the majority of webpages. The classifica-  
149 tion scheme that gave the initial impetus to research presented in this chapter was  
150 proposed by John Sinclair, first in the context of the EAGLES guidelines [9, 26].  
151 Among other dimensions of text classification Sinclair referred to the following six  
152 “intended outcomes of text production”:

- 153 1. information – reference compendia (Sinclair adds the following comment “an  
154 unlikely outcome, because texts are very rarely created merely for this purpose”);
- 155 2. discussion – polemic, position statements, argument;
- 156 3. recommendation – reports, advice, legal and regulatory documents;
- 157 4. recreation – fiction and non-fiction (biography, autobiography, etc.)
- 158 5. religion – holy books, prayer books, Order of Service (this label does not refer  
159 to religion as a topic);
- 160 6. instruction – academic works, textbooks, practical books.

161  
162 The typology is compact and applicable to webpages: only six top-level categories,  
163 each of which represents a variety of webpages, e.g., a page from Wikipedia  
164 is aimed at informing, a forum – at discussing, etc.

165 However, an attempt to apply these classes to the Web without any modification  
166 results in several problems. First, the boundary between look’n’feel and commu-  
167 nicative intentions is fuzzy. What is the reason for classifying a text as “recommen-  
168 dation”? Is this because it recommends an action or because it is classified as a  
169 report? A proposal issued by a think-tank of a political party can have “report” in its  
170 title, but in terms of its function it is very similar to a position statement published  
171 in a newspaper. The title of a publication is not the only reason for classifying it  
172 functionally, but in [9] no basis is given for classifying intentions.

173 Second, a functional classification assumes a certain degree of correlation  
174 between the function of a text and the language used to express this function. The  
175 function is *not defined* by linguistic features of respective texts, as otherwise the  
176 definition of genres depends on accidental features we choose to represent the genre,  
177 whereas its function in the society should be immune to such superficial variation.  
178 For instance, if narrative texts are defined by the number of past tense verbs [6],  
179 then narrative texts do not exist in Chinese, in which verbs do not have tenses. There  
180 might be a correlation between Chinese narrative texts and the amount of aspectual

## 7 In the Garden and in the Jungle

181 particles (e.g., *le*, *zhe*) or temporal adverbs (e.g., *zuotian*), but the dimension of nar-  
182 rativity has to be defined without relying on the features of an individual corpus. In  
183 other words, categories, such as narration, have to be specified taking into account  
184 the function a text has in the society, so that any comparison between corpora is  
185 made on the basis of categories more stable than linguistic features.<sup>3</sup> Nevertheless,  
186 it is reasonable to expect that texts contained in a single class of communicative  
187 aims (or “outcomes”) are more or less similar, e.g., narrative texts can be defined  
188 as texts reporting a sequence of events, and this correlates with certain linguistic  
189 features, which can be language- or even corpus-specific. On the other hand, if there  
190 is no similarity between regulatory documents and adverts (the latter are considered  
191 as a subclass of advice in Sinclair’s classification), there can be fewer reasons to  
192 keep them in the same class of “recommendations”. The same applies to joining  
193 academic works (such as the present chapter) and practical books (such as recipes)  
194 in the same category of “instructions”.

195 Third, decisions on document categorisation from their look’n’feel can be  
196 made by any reasonably confident user of those texts, while much more train-  
197 ing is needed to recognise more abstract functional categories. For instance,  
198 it is reasonable for the purposes of genre analysis to distinguish between  
199 blog entries aimed at discussion, news dissemination or recreation (entries  
200 with poetry or fiction), but naive annotators (much less ordinary Web-users)  
201 cannot make such distinctions reliably. In an experiment on webpage clean-  
202 ing [4], we attempted to annotate two sets of 60 webpages each in Chinese  
203 and English using a functional set of categories derived from Sinclair (the  
204 categories were advert, academic discussions, non-academic discussions,  
205 information, interview, instruction, fiction, news). Each page was anno-  
206 tated by two translation students who were familiar with classification of texts by  
207 their function and were given training to recognise the categories from this list.  
208 Nevertheless, the students failed to produce appropriate classification labels for  
209 some texts. Often both decisions made by the two annotators of the same text were  
210 different from the principles used in the typology suggested to them. For instance,  
211 a diary-like blog entry (<http://blogs.bootsnall.com/michelle/archives/006670.shtml>)  
212 was classified by one student as “information”, by another one as “news”, while  
213 it should have been classified as “non-academic discussions” along with all other  
214 private blog entries, if the instructions given as the basis for document categorisation  
215 were followed. This experiment suggests a gap between genre theory and the actual  
216 practice of average users.

217 Finally, some texts can be inherently ambiguous with respect to categories from  
218 Sinclair’s list. For instance, academic works are typically aimed at discussing states  
219 of affairs and making position statements; the boundary between “recommendation”  
220 and “discussion” is also frequently fuzzy. The same argument applies to traditional  
221

---

222  
223  
224 <sup>3</sup> This example assumes that the function of narration is actively used in the respective societies  
225 for approximately the same purposes, but for modern corpora this can be taken for granted.

226 rhetorical categories as well: the classes of descriptive, explicatory and argumenta-  
227 tive texts often overlap.

228 These considerations have led to the following adaptation of the original Sin-  
229 clair's typology:

- 230 1. *discussion* – all texts expressing positions and discussing a state of affairs
- 231 2. *information* – catalogues, glossaries, other lists (mostly containing incomplete  
232 sentences)
- 233 3. *instruction* – how-tos, FAQs, tutorials
- 234 4. *propaganda* – adverts, political pamphlets
- 235 5. *recreation* – fiction and popular lore
- 236 6. *regulations* – laws, small print, rules
- 237 7. *reporting* – newswires and informative broadcasts, police reports

238  
239 The present study is based on this typology, but I would refrain from saying that  
240 this is the final version. The category of *discussions* might need splitting, as it com-  
241 prises academic works and popular science, discussion forums and cases for support  
242 of academic projects, columns in newspapers and personal diaries, and so on. The  
243 difference between them can be described using other parameters of corpus classi-  
244 fication, such as the audience (professional or layman), publication medium (news-  
245 papers, forums, blogs), authorship (e.g., single or corporate). A multidimensional  
246 classification of this sort is more complex than a flat list of microgenres. However,  
247 the reason for this complexity is that many microgenres actually contain diverse  
248 text types. For instance, the category of blogs (frequently studied as a microgenre)  
249 does not define its functional content. Blogs are often studied from what is retrieved  
250 from a blogging website, like [blogspot.com](http://blogspot.com), which by itself only provides a tool  
251 that can help in publishing a chronologically ordered sequence of (short) texts. The  
252 genre is defined by the way this tool is used, e.g., to post newstems, publish fiction,  
253 discuss academic topics, or maintain personal diaries (with the two latter examples  
254 considered to be prototypical blogs). At the same time, a text can be published in a  
255 variety of possible publication media. For instance, a recipe (“instruction”) can be  
256 published in a blog entry, forum, newspaper or book.

257 Since the typology is meant to allow corpus comparison within and between  
258 languages, it should be complete: any webpage has to be classified according to a  
259 fixed number of predefined categories. Otherwise, it is difficult to compare corpora  
260 classified using different schemes. The functional principles for designing a typol-  
261 ogy mean that it is robust with respect to new emerging genres, as long as new  
262 communicative intentions do not emerge with new genres.

263 In designing a genre typology one open question is whether the typology is  
264 specific to an individual corpus, language or culture. Do we expect to use another  
265 typology to work with a corpus collected using different tools? Does the typology of  
266 English webpages apply to German, Russian or Chinese ones? The version proposed  
267 above corresponds to the mildest case of a culture-specific typology. It assumes  
268 that we derive the values of categories empirically from text categories which are  
269 more frequent in on the Web (across languages we are working with), also tak-  
270 ing into account the typology used in traditional reference corpora. “Mild” cultural

## 7 In the Garden and in the Jungle

271 dependence of the proposed typology means that it is specific to the current gen-  
272 eration of Web-derived corpora for languages with well-developed Internet culture.  
273 The typology listed above was developed from my attempts to classify English,  
274 German, Russian and Chinese webpages in my Internet corpora [24]. Most probably,  
275 it can be applied to describing the majority of modern webpages in, say, Arabic or  
276 Tagalog, while it may lack categories important for describing many texts written  
277 in the eighteenth century or in languages without an existing Internet culture like  
278 Brahui or Yukaghir, which might use the Web for purposes different from major  
279 languages.

280 Another open question concerns the ambiguity. One of the aims of the typology  
281 presented above is to reduce the ambiguity in comparison to the original Sinclair's  
282 classification, e.g., by splitting recommendation or adding a new category of report-  
283 ing. However, the ambiguity is wide-spread in real texts. This also concerns their  
284 communicative aims, so we can consider the possibility of using multiple labels,  
285 but the results of comparing two corpora with multiple labels are more difficult to  
286 interpret numerically. Therefore, in the study below each document gets a single  
287 label.

### 7.3 An Experiment in Automatic Classification of the Web

292 Once we have a typology, the next task is to classify I-EN and I-RU automati-  
293 cally and to compare their composition against traditional corpora (BNC and RNC  
294 respectively). A by-product of this study is the validation of the typology by check-  
295 ing whether its categories can be detected reliably and what confusion arises. One  
296 problem in this analysis is that supervised machine learning needs a large number of  
297 training examples, which are difficult to obtain from unclassified Web-derived cor-  
298 pora. Also, a comparison of I-EN and I-RU to their traditional counterparts implies  
299 classification of traditional corpora according to the same set of categories, while  
300 each corpus is documented using its own classification schemes.

301 Some genre labels used in BNC and RNC can be mapped to the more general  
302 functional categories listed above. For instance, academic (`W_ac_*`) and non-  
303 academic (`W_nonac_*`) papers from the BNC can be treated as "discussions",  
304 fiction and popular biographies as "recreational" texts, "propaganda" in the BNC  
305 is represented by `W_advert`. Not all genre labels can be mapped unambiguously,  
306 e.g., `W_commerce` or `W_email`. In addition to this, newspaper files in the BNC fre-  
307 quently consist of an entire issue and they contain a combination of genres, so they  
308 cannot be used for training purposes. Thus, the training corpus is a subset of the  
309 BNC.

310 This unambiguous mapping results in a "crisp" training corpus, which consists of  
311 texts definitely within the boundaries of the respective categories. For instance, we  
312 can populate the "instructions" category with texts marked as `W_instructional`  
313 in the BNC, 15 texts in total, such as recipe books, software manuals or DIY  
314 magazines. A clearer separation between text types is beneficial for the accuracy  
315 of cross-validation using the training corpus, but this eliminates other members

316 of this category, which do not have unambiguous labels in the BNC, e.g., text-  
317 books or academic tutorials. If we apply the model trained on a “crisp” corpus  
318 to the rest of the BNC, there is little chance that such texts will be recognised as  
319 “instructions”. On the other hand, including texts not explicitly labelled as such  
320 in the BNC, e.g., texts having “textbook” in their title or keywords, results in a  
321 “fuzzy” training corpus, which has a better coverage for each individual category,  
322 but contains more ambiguity, which might adversely affect the accuracy of the  
323 classifier.

324 The second problem with crisp corpora is that some BNC genre categories are  
325 easier to convert to corresponding communicative aims than others, so the training  
326 corpus can get significantly more discussions and recreational texts than other text  
327 types, e.g., 514 text can be classified as “recreation” vs. only 15 as “instruction”.  
328 The lack of balance can cause problems to machine learning algorithms, which pay  
329 attention to the probability of a category in the training corpus. In the end for instruc-  
330 tions and reporting categories I produced two versions, one was “crisp”, including,  
331 respectively, only `W_instructional` and `W_newsscript` texts. The other one was  
332 “fuzzy”, also including texts containing the word `textbook` in the title or keywords  
333 and `W_*_reportage` in its genre definition or news in the keywords. At the same  
334 time, the number of more frequent categories in the “fuzzy” corpus was reduced by  
335 random selection. Also, neither of the two corpora contains the category of “infor-  
336 mation”, as such texts (e.g., dictionaries or catalogue descriptions) have not been  
337 included in the BNC at all.

338 These subsets from traditional corpora were used to train SVM classifiers using  
339 the default parameters of Weka’s implementation of SVM [28]. Then, the models  
340 trained on a portion of traditional corpora were applied to the whole set. The features  
341 used for training were based on the frequency of POS trigrams describing individual  
342 texts, and also on the frequency of punctuation marks, e.g., quotes, exclamation and  
343 question marks each contributed to a feature. Given that the number of possible POS  
344 trigrams is fairly large resulting in a very sparse feature set, the study used the most  
345 significant POS trigrams selected using the Information Gain method, resulting in  
346 593 features for English and 577 features for Russian (the accuracy on a subset  
347 actually improves by a few percentage points in comparison to the full feature set  
348 and the resulting model is much faster).

349 In principle, web-related parameters can be additionally used to describe web-  
350 pages, such as the properties of originating URLs (e.g., the presence of `cgi-bin`  
351 or `~`), HTML tags (the use of fonts, tables or Javascript), navigation (links to other  
352 pages or links within a page), cf. [1, 20]. However, some information (such as  
353 HTML tags) has been lost in the process of corpus creation, and, more importantly,  
354 the chosen combination of POS trigrams with punctuation marks is applicable to  
355 both traditional written texts and webpages.

356 Table 7.1 compares the result of training using a “crisp” corpus against a “fuzzy”  
357 corpus. The accuracy is defined in Weka as the number of correctly classified  
358 instances (true positives) in the test corpus divided by its total size (averaged after  
359 10-fold cross-validation). As we can see the overall accuracy can be very high (up to  
360 97% with the crisp corpus), but this goes at the expense of the accuracy of assigning

## 7 In the Garden and in the Jungle

**Table 7.1** Comparing confusion matrices in training corpora

	a	b	c	d	e	f	←	Classified as		a	b	c	d	e	f	←	Classified as
361	194	1	6	6	1	0		a = Discussion	244	26	2	4	0	13		a = Discussion	
362	0	14	1	0	0	0		b = Instruction	19	49	3	4	1	0		b = Instruction	
363	5	1	47	1	0	0		c = Propaganda	10	3	46	1	0	0		c = Propaganda	
364	5	0	0	507	1	0		d = Recreation	3	1	0	194	0	1		d = Recreation	
365	0	1	0	0	76	0		e = Regulation	2	0	0	0	78	0		e = Regulation	
366	2	0	0	0	0	20		f = Reporting	14	0	0	0	0	29		f = Reporting	
367	Crisp BNC corpus (accuracy: 97%)								Fuzzy BNC corpus (accuracy: 86%)								
368																	
369																	
370																	
371																	
372																	
373																	
374																	
375																	
376																	
377																	
378																	
379																	
380																	
381																	
382																	
383																	
384																	
385																	
386																	
387																	
388																	
389																	
390																	
391																	
392																	
393																	
394																	
395																	
396																	
397																	
398																	
399																	
400																	
401																	
402																	
403																	
404																	
405																	

categories to examples outside clear-cut categories, when the classifier is applied to the rest of the BNC.<sup>4</sup> For instance, text A60, an introduction to international marketing, classified as *W\_commerce* in the BNC, is classified as “regulation” using the crisp training corpus, while it gets reclassified as “instruction” using the “fuzzy” one. This text does include formally written sentences that make it look like a piece of regulation (*International marketing is treated as a generic term covering the distinctions made in describing marketing activities as “international” or “multi-national” or “global”*), but the text as a whole is a textbook from the Kingston Business School. As a result, the crisp classifier treats only 86 texts in the whole BNC as “instructions”, while the fuzzy one finds 829 texts in this category, including A06 (a guide to becoming an actor), A0M (a karate handbook), A17 (a dog care magazine), none of which is treated as an instructional text in the BNC classification. Out of a random sample of 20 BNC texts automatically classified as “instructions”, only three texts should not belong to this category: C8X (poetry), KBS (a recorded dialogue) and KM4 (a recording from a business meeting). The results reported below are based on fuzzy training corpora.

<sup>4</sup> A similar pattern is evident in the accuracy drop from about 90% in the “crisp” 7-webgenre corpus to 66% in a fuzzy KI-04 corpus in experiments described in [22].

406 For English the procedure achieved the accuracy of 86% with 10-fold cross-  
407 validation, while the accuracy for Russian is significantly lower (74%), which can  
408 possibly be explained by the free word order, as well as by the greater number of  
409 morphological categories. For instance, the tagset used for English contains just  
410 four categories for nouns (common vs. proper, singular vs. plural), while in Russian  
411 nouns are described in terms of their number, gender, case, animacy, generating  
412 92 categories actually occurring in the training corpus. These factors make POS  
413 trigram statistics sparser, especially on the RNC texts, which are generally shorter  
414 than their BNC counterparts. At the same time, the greater granularity of POS cat-  
415 egories can help in distinguishing between genres. For instance, imperatives are a  
416 good indicator of instructions and propaganda, but in the English tagset such uses  
417 are treated identically to other base forms (infinitives and present simple forms).  
418 The same problem occurs with modal verbs: even if their functions are different  
419 and some modals are characteristic for specific genres (e.g., *shall* vs. *must*), in POS  
420 trigrams they are represented by a single tag.

421 Finally, the jungle of the Web was treated as being similar to the English gar-  
422 den, i.e., the models trained on the BNC and RNC were applied to English and  
423 Russian texts from the Internet corpora. First, the BNC and RNC models were  
424 applied to randomly selected subsets of 250 webpages from, respectively, I-EN  
425 and I-RU. The accuracy dropped considerably (down to 52% for English, 63% for  
426 Russian), but this gave the basis for creating a manually corrected training set to  
427 classify the entire Internet corpus. The drop in accuracy can be attributed to three  
428 factors<sup>5</sup>:

- 429
- 430 ● the balance of genres even in the fuzzy training corpus is quite different from  
431 what we have in the testing corpus: some classes are under-represented (report-  
432 ing), others are over-represented (fiction) or not represented in traditional corpora  
433 at all (information).
- 434 ● the Internet corpora are dirty in the sense that they contain some elements from  
435 original webpages not presented in the traditional corpora, such as navigation  
436 frames, ASCII art, standard headers. In spite of the best efforts to remove this noise,  
437 the accuracy of automatic cleaning is below 75% [4].
- 438 ● the language of the Internet is to some extent different from the language used  
439 in traditional corpora, e.g., not only British English is included in the annotated  
440 genre sample, FAQs are organised differently from tutorials listed in the BNC,  
441 the core of BNC texts stems from 1980s (the accuracy on the Russian sample  
442 was higher because the RNC is based on more recent texts, while I-RU is much  
443 more homogeneous in terms of the dialects it contains).
- 444
- 445
- 446
- 447

---

448 <sup>5</sup> The BNC has been retagged with TreeTagger, the same tool used for tagging I-EN, so there was  
449 no difference in the tagset and tagging between the two corpora (this could have caused variations  
450 in accuracy otherwise).

## 7 In the Garden and in the Jungle

## 7.4 Analysis of Results

The results of the automatic assessment of the composition of traditional and Internet corpora are presented in Table 7.2. The composition of the entire BNC and RNC was assessed by applying classifiers trained on their fuzzy subsets to their full content (BNC/F and RNC/F columns). I-EN and I-RU were assessed by their manually classified subsets of 250 texts each (I-EN/S and I-RU/S columns), and by applying classifiers trained on these subsets to their full content (I-EN/F and I-RU/F). Finally, the composition of ukWac, another corpus of English collected by crawling websites in the .uk domain, was also assessed by the same method (ukWac/F). To avoid data sparsity for classifiers, only texts longer than 300 words were used (this covers almost all texts in the BNC and more than 80% of I-EN and I-RU, 63% of ukWac).

**Table 7.2** Automatic assessment of corpus composition

Categories	BNC/F (%)	I-EN/S (%)	I-EN/F (%)	ukWac/F (%)	RNC/F (%)	I-RU/S (%)	I-RU/F (%)
Discussion	37.42	37.20	52.49	38.21	62.99	44.00	55.12
Information	0.00	6.00	4.03	5.03	0.00	0.40	0.06
Instruction	26.66	23.20	20.51	18.77	0.99	12.40	6.88
Propaganda	5.45	12.00	11.24	15.66	11.69	4.80	0.17
Recreation	21.43	4.00	0.97	1.03	14.17	24.80	27.46
Regulation	3.05	6.40	2.21	3.03	4.93	0.40	0.07
Reporting	6.00	11.20	8.54	18.27	5.22	13.20	10.24

## 7.4.1 Qualitative Assessment of Texts in Each Category

## 7.4.1.1 Discussion

This is the biggest category with a variety of subtypes. Automatic classifiers in general tended to overestimate the membership for this category, i.e., /F columns list more members than corresponding /S columns (especially for Russian). Texts classified in this way mostly include academic and newspaper articles (texts written for the professional audience vs. for the general public), as well as discussion forums and archived mailing lists.

## 7.4.1.2 Information

This macrogenre was not well represented in traditional corpora, such as the BNC and RNC, since corpus compilers tend to select running texts rather than catalogues or dictionaries. The procedure for collecting I-EN and I-RU also favoured running texts against incomplete descriptions by constructing longer queries, cf. [24, Section 2.2]. However, this macrogenre is common on the web. Pages classified as information include lists of people, places, businesses, objects, news stories, etc. A

496 fair amount of such texts (amounting to 15) managed to get into the random sample  
497 for English, even though fewer texts of this sort were detected in the full content  
498 of I-EN. There was only one text of this type in the Russian sample, which was  
499 not enough for training reliable classifiers. On the other hand, this macrogenre is  
500 more common in ukWac (which was produced by crawling, not by querying search  
501 engines). Texts of this type are important not only because of their amount, but also  
502 because of their potential to mislead POS taggers or other NLP tools. They often  
503 contain incomplete sentences with the visual boundary between their chunks often  
504 lost in the process of creating a plain text corpus.

505

#### 506 **7.4.1.3 Instruction**

507

508 The majority of texts classified with this label belong to two types:

509

- 510 • structured lists, such as FAQs, recipes, steps for assembling, repairing or main-  
511 taining something;
- 512 • advice written in a more narrative style, such as a recommendations, tutori-  
513 als, as well as some research papers, e.g., [http://www.privcom.gc.ca/media/nrc/  
514 opinion\\_021122\\_lf\\_e.asp](http://www.privcom.gc.ca/media/nrc/opinion_021122_lf_e.asp)

515 Such texts constitute about one quarter of either I-EN or ukWac, making it the  
516 second most frequent text type. However, it is found to be much less common in  
517 I-RU, though it is less common in the RNC as well. One possible reason for the  
518 apparent scarcity of such texts (they do constitute 12% of the sample from the Web)  
519 is the greater difficulty of detecting them in Russian. According to the Russian  
520 confusion matrix in Table 7.1, the majority of texts classified as “instruction” in  
521 the training set were classified as “discussion” by the automatic classifier. More  
522 research is needed to find features that can detect this class in Russian reliably.

523

#### 524 **7.4.1.4 Propaganda**

525

526 The amount of texts with propaganda of various sorts is in the range of 11% in I-EN  
527 to 16% in ukWac, while it is much less common in the BNC (5.5%). Pages classi-  
528 fied as propaganda typically promote goods and services, e.g., [http://www.hawaii-  
529 relocation.com/](http://www.hawaii-relocation.com/), which is not strictly speaking spam; this speaks against the reputa-  
530 tion of spam as the main polluter of Web-derived data.

531

#### 532 **7.4.1.5 Recreation**

533

534 It is known from other studies [24] that texts written with the purpose of  
535 recreation, such as fiction, are rare on the English Web (because of copy-  
536 right restrictions), while they are quite frequent for Russian. The present exper-  
537 iment confirms this to a certain extent. Nevertheless, such texts do exist in the  
538 two English Internet corpora. The most common microgenres are science fic-  
539 tion (often published under a Creative Commons license), collections of jokes  
540 (without explicit authorship), as well as all sorts of out-of-copyright fiction. The

## 7 In the Garden and in the Jungle

541 automatic classifier is also quite generous in assigning this category to texts,  
542 e.g., [http://42.blogs.warnock.me.uk/2006/05/cycling\\_fame.html](http://42.blogs.warnock.me.uk/2006/05/cycling_fame.html), that describes an  
543 event and is written in a chatty style (descriptions of events are normally classified  
544 as “reporting” otherwise). Anyway, one can argue that it is reasonable to classify  
545 texts of this sort as aimed at recreational reading.

546

### 547 7.4.1.6 Regulation

548

549 Texts classified in this way correspond to various rules, laws or official agreements,  
550 e.g., <http://contracts.onecle.com/talk/walsh.nso.2000.08.07.shtml>. According to the con-  
551 fusion matrix in Table 7.1 their detection in English is easy for the SVM classifier, so  
552 the figure for English in Table 7.2 can be assumed to be reliable. As for the Russian  
553 corpus, there was only one text of this type in the manually annotated sample, hence  
554 the classifier cannot be trained reliably. As a result there are numerous texts in I-RU  
555 automatically classified as “discussion”, while they can be reasonably treated as  
556 regulatory documents, e.g., <http://www.dmpmos.ru/law.asp?id=30020>.

557

### 558 7.4.1.7 Reporting

559

560 This category looks pretty uncontroversial. The original idea was to apply it to  
561 any type of newswires or reports about an event. Hence, the original classifier was  
562 trained on news scripts and reportage texts from the BNC (given the absence of  
563 police reports there). However, its application to webpages has identified other texts  
564 that can be reasonably treated as “reporting”, such as CVs, timelines of historic  
565 events or factual travel guides.

566

## 567 7.4.2 Assessing the Composition of ukWac

568

569 In this study I did not have time to evaluate the accuracy of genre assessment in  
570 ukWac on the basis of a large sample (around 250 documents). However, an initial  
571 estimate on transferring the classifiers trained on an I-EN sample to a new corpus  
572 can be made. Table 7.3 lists genres automatically assigned to documents collected  
573 from one website devoted to a large international conference. The results of clas-  
574 sification in all cases seem to be reasonable. For instance, the rules for taking part  
575 in a competition are treated as “instruction”, texts about exhibitors, sponsors and  
576 possibilities for advertising are treated as “propaganda”, while the conference pro-  
577 gramme has been classified as “reporting”.

578

579 However, several pages reasonably belonging to the same category are classified  
580 differently. Three issues of the newsletter are classified as “propaganda”, while the  
581 fourth one – as “discussion”. Out of the seven CVs of conference speakers (the last  
582 one combines CVs of several panelists), three are treated as “reporting”, while the  
583 other four – as “discussion”. There are inherent reasons for the differences in their  
584 automatic classification. The first three newsletters promoted the conference or its  
585 sponsors, while the last one mostly consisted of an informative interview. The CVs

**Table 7.3** Assessing genres in ukWac

586		
587	<a href="http://06.economie.co.uk/comp/rules.htm">http://06.economie.co.uk/comp/rules.htm</a>	Instruction
588	<a href="http://06.economie.co.uk/exhibitors/index.htm">http://06.economie.co.uk/exhibitors/index.htm</a>	Propaganda
589	<a href="http://06.economie.co.uk/location.htm">http://06.economie.co.uk/location.htm</a>	Discussion
590	<a href="http://06.economie.co.uk/newsletters/april2006.htm">http://06.economie.co.uk/newsletters/april2006.htm</a>	Propaganda
591	<a href="http://06.economie.co.uk/newsletters/aug1506.htm">http://06.economie.co.uk/newsletters/aug1506.htm</a>	Propaganda
592	<a href="http://06.economie.co.uk/newsletters/aug2806.htm">http://06.economie.co.uk/newsletters/aug2806.htm</a>	Propaganda
593	<a href="http://06.economie.co.uk/newsletters/may2006.htm">http://06.economie.co.uk/newsletters/may2006.htm</a>	Discussion
594	<a href="http://06.economie.co.uk/prog.htm">http://06.economie.co.uk/prog.htm</a>	Reporting
595	<a href="http://06.economie.co.uk/quiz.htm">http://06.economie.co.uk/quiz.htm</a>	Instruction
596	<a href="http://06.economie.co.uk/speakers/amy_domini.htm">http://06.economie.co.uk/speakers/amy_domini.htm</a>	Discussion
597	<a href="http://06.economie.co.uk/speakers/brian_spence.htm">http://06.economie.co.uk/speakers/brian_spence.htm</a>	Reporting
598	<a href="http://06.economie.co.uk/speakers/colin_baines.htm">http://06.economie.co.uk/speakers/colin_baines.htm</a>	Discussion
599	<a href="http://06.economie.co.uk/speakers/deborah_doane.htm">http://06.economie.co.uk/speakers/deborah_doane.htm</a>	Discussion
600	<a href="http://06.economie.co.uk/speakers/john_renesch.htm">http://06.economie.co.uk/speakers/john_renesch.htm</a>	Discussion
601	<a href="http://06.economie.co.uk/speakers/noreena_hertz.htm">http://06.economie.co.uk/speakers/noreena_hertz.htm</a>	Reporting
602	<a href="http://06.economie.co.uk/speakers/openforum.htm">http://06.economie.co.uk/speakers/openforum.htm</a>	Reporting
603	<a href="http://06.economie.co.uk/spons/additional.htm">http://06.economie.co.uk/spons/additional.htm</a>	Propaganda
604	<a href="http://06.economie.co.uk/spons/bursary.htm">http://06.economie.co.uk/spons/bursary.htm</a>	Propaganda
605	<a href="http://06.economie.co.uk/spons/index.htm">http://06.economie.co.uk/spons/index.htm</a>	Propaganda
606	<a href="http://06.economie.co.uk/spons/major.htm">http://06.economie.co.uk/spons/major.htm</a>	Propaganda
607	<a href="http://06.economie.co.uk/spons/opportunities.htm">http://06.economie.co.uk/spons/opportunities.htm</a>	Propaganda

in question were written in two different styles. One style describes the history of appointments (*Mike Kelly is Head of KPMG UK's Corporate Social Responsibility function. In 2002, Mike led KPMG's review of Environmental Risk Management at Morgan Stanley. Prior to coming to KPMG he was . . .*), while the other one emphasises the viewpoint of a person (*Variously described as a "business visionary" and as "a beacon lighting the way to a new paradigm", John Renesch stimulates people to think differently about work, leadership and the future. He believes that . . .*). The difference between these styles is obvious, but the decision made in each case is in the eye of the annotator (or automatic classifier), as views of the first person are described in his CV, even if they are less prominent than his function, while biographical details are also present in the second CV. The same argument applies to the difference between discussion and propaganda in the newsletters: the interview is informative, but it still promotes the company of the individual giving the interview.

## 7.5 Conclusions and Future Research

This chapter reports the first study, which was aimed at uncovering the genre composition of the entire jungle of the Web. The typology useful for classifying the entirety of webpages is still fluid. The main point of this study is to show that it is possible to estimate the composition of a corpus collected from the Web, even if it is a large corpus like I-EN (160 million words) or ukWac (2 billion words).

## 7 In the Garden and in the Jungle

631 In short the proposed procedure looks like this:

- 632 1. take a corpus with known composition (source corpus);
- 633 2. train a classifier on a subset;
- 634 3. apply it to a sample of a corpus with unknown composition (target corpus);
- 635 4. correct the sample and train a new classifier;
- 636 5. apply the new classifier to the rest of the corpus.

637  
638 If the system of genres used to describe the source corpus is identical to the  
639 genres needed to assess the target corpus, the whole source corpus can be used in  
640 Step 2. In another experiment, I classified I-EN and ukWac using the entire set of  
641 70 genres of the BNC and four main genre categories of the Brown corpus (press,  
642 fiction, nonfiction and misc), following the results reported in [25]. This gives us  
643 data for comparing genre composition of a variety of corpora or for selecting sub-  
644 sets to study them more closely. For instance, 18,715 webpages in ukWac have been  
645 classified as `personal_letters` using the BNC-trained classifier, with the vast  
646 majority of them being diary entries coming from blogs. So this classifier provides  
647 a useful mechanism for finding and studying diary-like blogs. However, the value  
648 of such tests is limited, as the experiments with the BNC and RNC (Section 7.3)  
649 show that the process of retraining using a subset of the target corpus (Steps 3  
650 and 4) is necessary to improve the accuracy of the classifier on data from the target  
651 corpus.

652 Even the results for the validated classifiers have to be taken *cum grano salis*.  
653 It is tempting to refer to the results in Table 7.2 as saying that the composition  
654 of the Web is as follows: instructions – one quarter, advertising and propaganda –  
655 10–15%, lists and catalogues – 5%, regulations – about 3%, etc. However, there are  
656 obvious limitations on extrapolating this study. First, the results are based on I-EN  
657 and ukWac, Web-derived corpora collected in a particular way. Both corpora contain  
658 only HTML pages (PDF files or Word documents were not used); the procedure for  
659 their collection favoured finding examples of running text at the expense of “index”  
660 pages or other collections of links (even though the methods for rejecting such pages  
661 were specific to each corpus), duplicate webpages in both corpora were discarded.  
662 Other methods of corpus collection might favour other slices of the Web and get  
663 different results.

664 Second, my training corpora used in Step 4 consisted of 250 webpages. This led  
665 to a limited number of training examples for less frequent categories. For instance,  
666 the Russian training sample contained just one example of texts classified as “infor-  
667 mation” and “regulation”, respectively. This is indicative of the fact that these text  
668 types are not very frequent in the rest of I-RU, see the discussion of sampling statis-  
669 tics in [24], but single examples do not give sufficient information for classifying  
670 unseen texts of this type. Some other macrogenres have more training examples, but  
671 they are still represented by a small number of microgenres. For instance, out of 16  
672 texts classified as “regulation” in the English sample, there was no text belonging  
673 to the microgenre of “contractual agreements”, e.g., *Either party shall be entitled*  
674 *on written notice to terminate . . .* Thus, texts of this type from the full corpus are  
675 less likely to be classified as regulations. This suggests the need to have a greater

676 variety of texts in the training corpus, even at the expense of random selection of the  
677 sample, cf. the discussion about a representative corpus of webgenres in Chapter 5  
678 by Santini's, this book.

679 The features discriminating between genres in the experiments described above  
680 were based on POS trigrams and punctuation statistics. However, more research  
681 is needed into detection of reliable genre indicators, including lexical features  
682 (e.g., keywords,<sup>6</sup> frequency bands, n-grams, lexical density, etc), grammatical fea-  
683 tures other than POS trigrams (the latter are quite sparse in morphologically rich  
684 languages, such as Russian), text statistics (average document or sentence length,  
685 web-specific markup statistics or URL components, etc). More research is also  
686 needed into methods for more efficient population of the feature set with features  
687 corresponding to individual categories.

688 A more general remark concerns the merits of using macrogenres (such as used  
689 in this study) vs. microgenres. As mentioned above, the use of the seven macrogenre  
690 categories studied in this chapter results in a very coarse classification. If our task  
691 to study the microgenre of prototypical blogs, i.e., short personal notes published  
692 in a chronological order, the results reported in Section 7.3 are of little help, as this  
693 microgenre is contained within in a much bigger macrogenre of "discussions". In  
694 addition to this, macrogenre categories are usually abstract, so their reliable recog-  
695 nition requires training. Unlike "look-n-feel" categories, ordinary Internet users or  
696 people outside of the community of genre scholars can find it difficult to use them,  
697 e.g., for refining the results of web searches.

698 However, we need a common yardstick for describing the composition of corpora  
699 collected using different methods from different sources, so that we can compare  
700 the proportion of genres in the BNC and ukWac, or in ukWac and deWac. Table 7.2  
701 demonstrates the possibility of achieving this using a compact genre typology. A list  
702 of 70 genres of the BNC or 78 webgenres suggested in [21] would be more difficult  
703 to apply as a yardstick because of various reasons:

- 704 • the ambiguity usually increases with the number of categories, e.g., Wikipedia  
705 entries are (unintentionally) mentioned as an example in the categories of "Ency-  
706 clopedias" and "Feature stories" in [21];
- 707 • the accuracy of automatic classification usually drops if the classifier has to  
708 distinguish between a larger number of possible choices, e.g., the F-measure  
709 reported in [27] is about 50% for 20 genres vs. 80% in Table 7.2, while machine  
710 learning methods used in the two studies are very similar;
- 711 • it is difficult to analyse results described in terms of a large number of different  
712 parameters (even the seven categories in Table 7.2 present problems for interpre-  
713 tation; if Table 7.2 was expanded to 78 categories, it would be almost impossible  
714 to interpret).

---

719  
720 <sup>6</sup> The use of keywords for genre detection has been studied, e.g., in [29] or [8].

## 7 In the Garden and in the Jungle

721 **Acknowledgments** I'm grateful to Silvia Bernardini, Adam Kilgarriff, Katja Markert and Marina  
 722 Santini for useful discussions. The usual disclaimers apply. The tools for genre classification  
 723 described in this chapter and the results of classifications of the Internet corpora are available  
 724 from <http://corpus.leeds.ac.uk/serge/webgenres/>

727 **References**

- 729 1. Allen, P., J.A. Bateman, and J.L. Delin. 1999. Genre and layout in multimodal docu-  
 730 ments: Towards an empirical account. In *Proceedings of the AAAI Fall Symposium on*  
 731 *Using Layout for the Generation, Understanding, or Retrieval of Documents*, eds. R.  
 732 Power and D. Scott, 27–34. Cape Cod, MA: American Association for Artificial Intelli-  
 733 gence. URL [http://www.fb10.uni-bremen.de/anglistik/langpro/projects/gem/downloads/allen-](http://www.fb10.uni-bremen.de/anglistik/langpro/projects/gem/downloads/allen-bateman-delin.PDF)  
 734 [bateman-delin.PDF](http://www.fb10.uni-bremen.de/anglistik/langpro/projects/gem/downloads/allen-bateman-delin.PDF)
- 735 2. Baroni, M., and S. Bernardini. 2004. Bootcat: Bootstrapping corpora and terms from the web.  
 736 In *Proceedings of LREC2004*. Lisbon.
- 737 3. Baroni, M., and A. Kilgarriff. 2006. Large linguistically-processed Web corpora for multiple  
 738 languages. In *Companion Volume to Proceedings of the European Association of Computa-*  
 739 *tional Linguistics*, 87–90. Trento.
- 740 4. Baroni, M., F. Chantree, A. Kilgarriff, and S. Sharoff. 2008. Cleaneval: A competition for  
 741 cleaning web pages. In *Proceedings of the 6th Language Resources and Evaluation Con-*  
 742 *ference, LREC 2008*. Marrakech. URL [http://corpus.leeds.ac.uk/serge/publications/lrec2008-](http://corpus.leeds.ac.uk/serge/publications/lrec2008-cleaneval.pdf)  
 743 [cleaneval.pdf](http://corpus.leeds.ac.uk/serge/publications/lrec2008-cleaneval.pdf)
- 744 5. Biber, D. 1988. *Variations across speech and writing*. Cambridge, MA: Cambridge Univer-  
 745 sity Press.
- 746 6. Biber, D., and J. Kurjian. 2006. Towards a taxonomy of web registers and text types: A mul-  
 747 tidimensional analysis. In *Corpus linguistics and the web*, eds. M. Hundt, N. Nesselhauf, and  
 748 C. Biewer, 109–131. Amsterdam: Rodopi.
- 749 7. Braslavski, P. 2004. Document style recognition using shallow statistical analysis. In *ESSLLI*  
 750 *2004 Workshop on Combining Shallow and Deep Processing for NLP*, 1–9.
- 751 8. Crossley, S.A., and M. Lowerse. 2007. Multi-dimensional register classification using  
 752 bigrams. *International Journal of Corpus Linguistics* 12(4):453–478.
- 753 9. EAGLES. 1996. Preliminary recommendations on text typology. Technical Report EAG-  
 754 TCWG-TTYP/P, Expert Advisory Group on Language Engineering Standards document.  
 755 URL <http://www.ilc.cnr.it/EAGLES96/texttyp/texttyp.html>
- 756 10. Ferraresi, A. 2007. Building a very large corpus of English obtained by web crawling: ukwac.  
 757 Master's thesis, University of Bologna.
- 758 11. Halliday, M.A.K. 1985. *An introduction to functional grammar*. London: Edward Arnold.
- 759 12. Jakobson, R. 1960. Linguistics and poetics. In *Style in Language*, ed. T.A. Sebeok, 350–377.  
 760 Cambridge, MA: MIT Press.
- 761 13. Joho, H., and M. Sanderson. 2004. The SPIRIT collection: An overview of a large web col-  
 762 lection. *SIGIR Forum* 38(2):57–61. doi: <http://doi.acm.org/10.1145/1041394.1041395>
- 763 14. Kessler, B., Nunberg, G., and H. Schütze. 1997. Automatic detection of text genre. In *Pro-*  
 764 *ceedings of the 35th ACL/8th EACL*, 32–38.
- 765 15. Kilgarriff, A. 2001. The web as corpus. In *proceeding of corpus linguistics 2001*. Lancaster.  
 URL <http://www.itri.bton.ac.uk/techreports/ITRI-01-14.abs.html>
16. Lee, D. 2001. Genres, registers, text types, domains, and styles: Clarifying the concepts and  
 navigating a path through the BNC jungle. *Language Learning and Technology* 5(3):37–72.  
 URL <http://ilt.msu.edu/vol5num3/pdf/lee.pdf>
17. Macdonald, C., and I. Ounis. 2006. The TREC blogs06 collection: Creating and analysing a  
 blog test collection. Technical Report TR-2006-224, Department of Computing Science, Uni-  
 versity of Glasgow. URL <http://ir.dcs.gla.ac.uk/terrier/publications/macdonald06creating.pdf>

AQ2

AQ3

- 766 18. Martin, J.R. 1984. Language, register and genre. In *Children Writing: Reader (ECT language*  
767 *studies: Children writing)*, ed. F. Christie, 21–30. Geelong, VIC: Deakin University Press.
- 768 19. Mehler, A., and R. Gleim. 2006. The net for the graphs – towards webgenre representation for  
769 corpus linguistic studies. In *WaCky! Working papers on the Web as Corpus*, eds. M. Baroni  
770 and S. Bernardini. Bologna: Gedit.
- 771 20. Meyer zu Eissen, S., and B. Stein. 2004. Genre classification of web pages. In *Proceedings of*  
772 *the 27th German Conference on Artificial Intelligence*. Ulm.
- 773 21. Rehm, G., M. Santini, A. Mehler, P. Braslavski, R. Gleim, A. Stubbe, S. Symonenko, M.  
774 Tavosanis, and V. Vidulin. 2008. Towards a reference corpus of web genres for the evaluation  
775 of genre identification systems. In *Proceedings of the 6th Language Resources and Evaluation*  
776 *Conference, LREC 2008*. Marrakech.
- 777 22. Santini, M. 2007. Automatic identification of genre in web pages. PhD thesis, University of  
778 Brighton.
- 779 23. Sharoff, S. 2005. Methods and tools for development of the Russian reference corpus. In  
780 *Corpus linguistics around the world*, eds. D. Archer, A. Wilson, and P. Rayson, 167–180.  
781 Amsterdam: Rodopi.
- 782 24. Sharoff, S. 2006. Creating general-purpose corpora using automated search engine queries. In  
783 *WaCky! Working papers on the Web as Corpus*, eds. M. Baroni and S. Bernardini. Bologna:  
784 Gedit. <http://wackybook.sslmit.unibo.it>
- 785 25. Sharoff, S. 2007. Classifying web corpora into domain and genre using automatic feature  
786 identification. In *Proceedings of Web as Corpus Workshop*. Louvain-la-Neuve.
- 787 26. Sinclair, J. 2003. Corpora for lexicography. In *A practical guide to lexicography*, ed. P. van  
788 Sterkenberg, 167–178. Amsterdam: Benjamins.
- 789 27. Vidulin, V., M. Luštrek, and M. Gams. 2007. Using genres to improve search engines.  
790 In *Proceeding Towards Genre-Enabled Search Engines: The Impact of NLP*. RANLP,  
791 URL [http://dis.ijs.si/MitjaL/documents/Vidulin-Using\\_Genres\\_to\\_Improve\\_Search\\_Engines-](http://dis.ijs.si/MitjaL/documents/Vidulin-Using_Genres_to_Improve_Search_Engines-RANLP-07-TGESE.pdf)  
792 [RANLP-07-TGESE.pdf](http://dis.ijs.si/MitjaL/documents/Vidulin-Using_Genres_to_Improve_Search_Engines-RANLP-07-TGESE.pdf)
- 793 28. Witten, I.H., and E. Frank. 2005. *Data Mining: Practical machine learning tools and tech-*  
794 *niques*. San Francisco, CA: Morgan Kaufmann.
- 795 29. Xiao, Z., and A. McEnery. 2005. Three genres in modern American English. *Journal of*  
796 *English Linguistics* 33(1):62–82.
- 797
- 798
- 799
- 800
- 801
- 802
- 803
- 804
- 805
- 806
- 807
- 808
- 809
- 810

811 **Chapter 7**

812

---

813	<b>Query No.</b>	<b>Query</b>
814		
815	AQ1	Please provide keywords for Online version.
816	AQ2	Please provide location for the reference [7].
817	AQ3	Please provide location for the reference [14].
818		

---

819

820

821

822

823

824

825

826

827

828

829

830

831

832

833

834

835

836

837

838

839

840

841

842

843

844

845

846

847

848

849

850

851

852

853

854

855

UNCORRECTED PROOF