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, LS2 9JT Leeds, UK
e-mail: s.sharoff@leeds.ac.uk

A. Mehler et al. (eds.), *Genres on the Web*, Text, Speech and Language Technology 42, DOI 10.1007/978-90-481-9178-9_7,

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

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

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

[1] Throughout this chapter I refer to BNC texts using their ids from the BNC Index, which is available from http://clix.to/davidlee00

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

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

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

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

7.2 Text Typology for the Web

Approaches to classifying texts into genres can be grouped into two main classes. The first class identifies genres of documents on the basis of what can be called "look'n'feel" properties, e.g., FAQ, forum or recipe, while the second class detects broad functional classes, e.g., description or argumentation, cf. the discussion in [16] or [5].

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

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

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

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

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

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

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

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

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

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

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

[3] This example assumes that the function of narration is actively used in the respective societies for approximately the same purposes, but for modern corpora this can be taken for granted.

rhetorical categories as well: the classes of descriptive, explicatory and argumentative texts often overlap.

These considerations have led to the following adaptation of the original Sinclair's typology:

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

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

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

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

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

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

7.3 An Experiment in Automatic Classification of the Web

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

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

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

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

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

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

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

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

Table 7.1 Comparing confusion matrices in training corpora

a	b	c	d	e	f	⇐	Classified as
194	1	6	6	1	0	a =	Discussion
0	14	1	0	0	0	b =	Instruction
5	1	47	1	0	0	c =	Propaganda
5	0	0	507	1	0	d =	Recreation
0	1	0	0	76	0	e =	Regulation
2	0	0	0	0	20	f =	Reporting

Crisp BNC corpus (accuracy: 97%)

a	b	c	d	e	f	⇐	Classified as
244	26	2	4	0	13	a =	Discussion
19	49	3	4	1	0	b =	Instruction
10	3	46	1	0	0	c =	Propaganda
3	1	0	194	0	1	d =	Recreation
2	0	0	0	78	0	e =	Regulation
14	0	0	0	0	29	f =	Reporting

Fuzzy BNC corpus (accuracy: 86%)

TP rate	FP rate	Precision	Recall	F-measure	Class
0.869	0.102	0.843	0.869	0.856	Discussion
0.623	0.046	0.608	0.623	0.615	Instruction
0.783	0.010	0.870	0.783	0.825	Propaganda
0.975	0.016	0.956	0.975	0.965	Recreation
0.950	0.001	0.987	0.950	0.968	Regulation
0.605	0.016	0.703	0.605	0.650	Reporting
0.800	0.030	0.830	0.800	0.810	*Average*

Fuzzy BNC, detailed accuracy by class

a	b	c	d	e	f	⇐	Classified as
721	2	89	66	32	55	a =	Discussion
41	17	14	4	12	2	b =	Instruction
176	8	394	3	33	13	c =	Propaganda
51	2	2	890	0	4	d =	Recreation
55	12	45	0	339	19	e =	Regulation
101	3	23	18	23	183	f =	Reporting

Russian fuzzy training corpus (accuracy: 74%)

categories to examples outside clear-cut categories, when the classifier is applied to the rest of the BNC.[4] 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 "multinational" 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.

[4] 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].

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

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

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

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

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

7.4.1.3 Instruction

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

- structured lists, such as FAQs, recipes, steps for assembling, repairing or maintaining something;
- advice written in a more narrative style, such as a recommendations, tutorials, as well as some research papers, e.g., http://www.privcom.gc.ca/media/nr-c/opinion_021122_lf_e.asp

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

7.4.1.4 Propaganda

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

7.4.1.5 Recreation

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

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

7.4.1.6 Regulation

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

7.4.1.7 Reporting

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

7.4.2 Assessing the Composition of ukWac

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

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

Table 7.3 Assessing genres in ukWac

http://06.economie.co.uk/comp/rules.htm	Instruction
http://06.economie.co.uk/exhibitors/index.htm	Propaganda
http://06.economie.co.uk/location.htm	Discussion
http://06.economie.co.uk/newsletters/april2006.htm	Propaganda
http://06.economie.co.uk/newsletters/aug1506.htm	Propaganda
http://06.economie.co.uk/newsletters/aug2806.htm	Propaganda
http://06.economie.co.uk/newsletters/may2006.htm	Discussion
http://06.economie.co.uk/prog.htm	Reporting
http://06.economie.co.uk/quiz.htm	Instruction
http://06.economie.co.uk/speakers/amy_domini.htm	Discussion
http://06.economie.co.uk/speakers/brian_spence.htm	Reporting
http://06.economie.co.uk/speakers/colin_baines.htm	Discussion
http://06.economie.co.uk/speakers/deborah_doane.htm	Discussion
http://06.economie.co.uk/speakers/john_renesch.htm	Discussion
http://06.economie.co.uk/speakers/noreena_hertz.htm	Reporting
http://06.economie.co.uk/speakers/openforum.htm	Reporting
http://06.economie.co.uk/spons/additional.htm	Propaganda
http://06.economie.co.uk/spons/bursary.htm	Propaganda
http://06.economie.co.uk/spons/index.htm	Propaganda
http://06.economie.co.uk/spons/major.htm	Propaganda
http://06.economie.co.uk/spons/opportunities.htm	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).

In short the proposed procedure looks like this:

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

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

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

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

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

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

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

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

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

[6] The use of keywords for genre detection has been studied, e.g., in [29] or [8].

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

References

1. Allen, P., J.A. Bateman, and J.L. Delin. 1999. Genre and layout in multimodal documents: Towards an empirical account. In *Proceedings of the AAAI Fall Symposium on Using Layout for the Generation, Understanding, or Retrieval of Documents*, eds. R. Power and D. Scott, 27–34. Cape Cod, MA: American Association for Artificial Intelligence. URL http://www.fb10.uni-bremen.de/anglistik/langpro/projects/gem/downloads/allen-bateman-delin.PDF
2. Baroni, M., and S. Bernardini. 2004. Bootcat: Bootstrapping corpora and terms from the web. In *Proceedings of LREC2004*. Lisbon.
3. Baroni, M., and A. Kilgarriff. 2006. Large linguistically-processed Web corpora for multiple languages. In *Companion Volume to Proceedings of the European Association of Computational Linguistics*, 87–90. Trento.
4. Baroni, M., F. Chantree, A. Kilgarriff, and S. Sharoff. 2008. Cleaneval: A competition for cleaning web pages. In *Proceedings of the 6th Language Resources and Evaluation Conference, LREC 2008*. Marrakech. URL http://corpus.leeds.ac.uk/serge/publications/lrec2008-cleaneval.pdf
5. Biber, D. 1988. *Variations across speech and writing*. Cambridge, MA: Cambridge University Press.
6. Biber, D., and J. Kurjian. 2006. Towards a taxonomy of web registers and text types: A multidimensional analysis. In *Corpus linguistics and the web*, eds. M. Hundt, N. Nesselhauf, and C. Biewer, 109–131. Amsterdam: Rodopi.
7. Braslavski, P. 2004. Document style recognition using shallow statistical analysis. In *ESSLLI 2004 Workshop on Combining Shallow and Deep Processing for NLP*, 1–9. Nancy.
8. Crossley, S.A., and M. Lowerse. 2007. Multi-dimensional register classification using bigrams. *International Journal of Corpus Linguistics* 12(4):453–478.
9. EAGLES. 1996. Preliminary recommendations on text typology. Technical Report EAG-TCWG-TTYP/P, Expert Advisory Group on Language Engineering Standards document. URL http://www.ilc.cnr.it/EAGLES96/texttyp/texttyp.html
10. Ferraresi, A. 2007. Building a very large corpus of English obtained by web crawling: ukwac. Master's thesis, University of Bologna.
11. Halliday, M.A.K. 1985. *An introduction to functional grammar*. London: Edward Arnold.
12. Jakobson, R. 1960. Linguistics and poetics. In *Style in Language*, ed. T.A. Sebeok, 350–377. Cambridge, MA: MIT Press.
13. Joho, H., and M. Sanderson. 2004. The SPIRIT collection: An overview of a large web collection. *SIGIR Forum* 38(2):57–61. doi: http://doi.acm.org/10.1145/1041394.1041395
14. Kessler, B., Nunberg, G., and H. Schütze. 1997. Automatic detection of text genre. In *Proceedings of the 35th ACL/8th EACL*, 32–38. Madrid.
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://llt.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, University of Glasgow. URL http://ir.dcs.gla.ac.uk/terrier/publications/macdonald06creating.pdf

18. Martin, J.R. 1984. Language, register and genre. In *Children Writing: Reader (ECT language studies: Children writing)*, ed. F. Christie, 21–30. Geelong, VIC: Deakin University Press.
19. Mehler, A., and R. Gleim. 2006. The net for the graphs – towards webgenre representation for corpus linguistic studies. In *WaCky! Working papers on the Web as Corpus*, eds. M. Baroni and S. Bernardini. Bologna: Gedit.
20. Meyer zu Eissen, S., and B. Stein. 2004. Genre classification of web pages. In *Proceedings of the 27th German Conference on Artificial Intelligence*. Ulm.
21. Rehm, G., M. Santini, A. Mehler, P. Braslavski, R. Gleim, A. Stubbe, S. Symonenko, M. Tavosanis, and V. Vidulin. 2008. Towards a reference corpus of web genres for the evaluation of genre identification systems. In *Proceedings of the 6th Language Resources and Evaluation Conference, LREC 2008*. Marrakech.
22. Santini, M. 2007. Automatic identification of genre in web pages. PhD thesis, University of Brighton.
23. Sharoff, S. 2005. Methods and tools for development of the Russian reference corpus. In *Corpus linguistics around the world*, eds. D. Archer, A. Wilson, and P. Rayson, 167–180. Amsterdam: Rodopi.
24. Sharoff, S. 2006. Creating general-purpose corpora using automated search engine queries. In *WaCky! Working papers on the Web as Corpus*, eds. M. Baroni and S. Bernardini. Bologna: Gedit. http://wackybook.sslmit.unibo.it
25. Sharoff, S. 2007. Classifying web corpora into domain and genre using automatic feature identification. In *Proceedings of Web as Corpus Workshop*. Louvain-la-Neuve.
26. Sinclair, J. 2003. Corpora for lexicography. In *A practical guide to lexicography*, ed. P. van Sterkenberg, 167–178. Amsterdam: Benjamins.
27. Vidulin, V., M. Luštrek, and M. Gams. 2007. Using genres to improve search engines. In *Proceeding Towards Genre-Enabled Search Engines: The Impact of NLP*. RANLP, URL http://dis.ijs.si/MitjaL/documents/Vidulin-Using_Genres_to_Improve_Search_Engines-RANLP-07-TGESE.pdf
28. Witten, I.H., and E. Frank. 2005. *Data Mining: Practical machine learning tools and techniques*. San Francisco, CA: Morgan Kaufmann.
29. Xiao, Z., and A. McEnery. 2005. Three genres in modern American English. *Journal of English Linguistics* 33(1):62–82.