

In the garden and in the jungle: comparing genres in the BNC and Internet

Serge Sharoff

Centre for Translation Studies, University of Leeds

Abstract. In this paper I will present an approach to classifying the Web into genres. The goal is to have a compact system of categories that can be assigned with little ambiguity to almost every Internet page. The proposed typology is organised from the functional viewpoint: generalised categories for genre classification correspond to major aims of text production, such as ‘discussion’ or ‘instruction’. This paper compares the genre distributions in English and Russian Internet corpora against their human-collected counterparts (BNC and RNC) in terms of these classes using probabilistic classifiers.

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’ [1]. 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 [2]. Intuitively this claim is plausible: webpages present a much greater variety of text types and genres than the 4055 texts in the BNC classified into 70 genres. However, this does not release us from the duty of studying this jungle.

Nowadays it is relatively easy to collect a large corpus from the Web, either using search engines [3] or web crawlers [4,5], 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 [6] 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. Similarly, a single large text taken from a newspaper and classified as world affairs in the BNC can contain home and foreign news, social commentary, gossips, etc. Many genres also remain underdescribed. Even though there are textbooks in the BNC (for instance, texts EVW or GVS),

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.¹ 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 [3], SPIRIT [4] or deWaC [5].

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 follow traditions of printed media, e.g., on-line newspapers, while others are markedly different from any known printed counterpart, e.g., chat rooms. Third, given the huge number of pages in Internet corpora, e.g., more than 60,000 in I-EN [3], we need automatic methods that can identify genres reliably and be applicable to arbitrary webpages. 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 list discussions, project proposals, progress reports, minutes of meetings, job descriptions, etc. 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, whereas the function of blogs is similar to that of diaries or opinion columns in newspapers, and is different from them in the audience size, distribution mode and authorship.

One approach to studying genres on the Web is to start with a genre (or a group of genres), such as blogs [7] or conference websites [8]. Then we can analyse linguistic features specific to this genre and learn how to identify a text as belonging or not belonging to it. This approach focuses on individual characteristics of each genre. Another approach is aimed at saying “sensible and useful things about any text” that exists on the Web.² This approach can offer a very superficial description for many genres studied in the first approach,

¹ Data from the BNC Index <http://clix.to/davidlee00>

² The quote refers to the purposes Michael Halliday intended for his “Introduction to Functional Grammar” [9]

but without a possibility to use a compact text typology the first approach risks degenerating into an infinite list of genre types to account for all possible webpages.

In this paper I will outline an approach to text classification that can be used to describe the majority of texts on the Web using a small number (less than 10) of categories, so that we can broadly assess the composition of genres in a web-derived corpus, compare it against other Internet collections and traditional corpora, as well as against corpora in other languages (Section 2). Then (in Section 3) I will present an experiment for detection of text categories in traditional representative corpora against Internet corpora in English and Russian. 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 [10]. Finally, in Section 4 I will discuss the similarities and differences between Web corpora and their manually collected counterparts.

The study concerns English and Russian corpora collected from the Web using random queries to search engines [3]. Below these corpora are referred to as I-EN and I-RU. 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 collected from the Web. In the last section, I also report a small experiment of applying the same methodology to classifying ukWac, another English corpus collected by crawling websites in the .uk domain [11].

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 [1] or [12].

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 this label 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 does not give the flexibility to compare webcorpora to their traditional counterparts (Problem 5).

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 [13, Ch. 2]. Without giving an explicit text typology, James Martin defines genres via ‘staged, goal-oriented, purposeful activity in which speakers engage as members of our culture’ [14, p. 25]. In [15], 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 [16], the classification is done into five functional styles: fiction, journalism, official, academic, everyday language, following a tradition that stems from Jakobson [17].

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 paper was proposed by John Sinclair, first in the context of the EAGLES guidelines [18,19]. 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 does not refer to religion as a topic, which is the label used for referring to the domain of a text);
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 web pages, e.g., a page from Wikipedia is aimed at informing, a forum — at discussing, etc.

However, an attempt to apply these classes to Internet corpora 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 [18] 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, then narrative texts do not exist in Chinese, in which verbs do not have tenses. Nevertheless, it is reasonable to expect that texts contained in a single

class of aims (or ‘outcomes’) are more or less similar in terms of their features, e.g., narrative texts can be defined as text reporting a sequence of events, and this correlates with certain linguistic features. On the other hand, if there is no similarity between regulatory documents and adverts (the latter are considered as a subclass of advice), it is not reasonable to keep them in the same class of ‘recommendations’. The same problem applies to joining academic works (such as the present paper) and practical books (such as recipes) in the same category of ‘instructions’.

Third, decisions for classifying documents on the basis of 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 to distinguish between blog entries posted for the purpose of discussion, information or recreation (entries with poetry or fiction), but naive annotators (much less Internet users) cannot make such distinctions reliably. In an experiment on web page cleaning [20], we attempted to annotate two sets of 60 webpages each in Chinese and English respectively using a functional set of categories derived from Sinclair (advert, news, information, fiction, interview, instruction, academic discussions, non-academic discussions). Each page was annotated by two translation students who were aware of classification of texts by their function and were given training to recognise the categories from this list. Nevertheless, they failed to agree on the classification label for many texts. Often both decisions made by the annotators were different from our view as supervisors. 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 clearly represents the viewpoint of its author and, according to this classification, should be classified as ‘non-academic discussions’.

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 rhetorics as well: the categories 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, 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. A blog is just a tool that can help in publishing a chronologically ordered sequence of texts. The genre is defined by the way this tool is used, e.g., to post newstems, discuss parenting or academic topics, publish personal diaries or fiction. 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, forum, newspaper or book. At the same time, if our task to study the microgenre of prototypical blogs, i.e., short personal notes on a particular topic, the proposed typology is too coarse, as this microgenre is mixed with other types of discussions.

The intention of making it possible to compare corpora within and between languages also assumes that the typology is 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 Internet? 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 Internet corpora (across languages we are working with), also taking into account also the typology used in traditional reference corpora. "Mild" cultural dependence of the proposed typology means that it is specific to the current generation of Internet corpora for languages with well-developed Internet culture. The typology is aimed at describing any modern webpage in, say, Arabic or Tagalog, while it may lack categories important for describing many texts written in the 18th century or in languages without an existing Internet culture like Brahui or Yukaghir, which might use Internet 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.

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 validation of the typology by checking whether its categories can be detected reliably and what confusion can arise. One problem in this analysis is that supervised machine learning needs a large number of training examples, which are difficult to obtain from unclassified Internet corpora. Also, comparison of I-EN and I-RU to their traditional counterparts implies classification of traditional corpora according to the same set of categories, while they are documented using their own classification schemes.

Some genre labels used in BNC and RNC can be mapped to the more general 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 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 more clear 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 marked 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 adversely affects 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

Table 1. Comparing confusion matrices in training corpora

a	b	c	d	e	f	← classified as	a	b	c	d	e	f	← classified as
194	1	6	6	1	0	a = discussion	244	26	2	4	0	13	a = discussion
0	14	1	0	0	0	b = instruction	19	49	3	4	1	0	b = instruction
5	1	47	1	0	0	c = propaganda	10	3	46	1	0	0	c = propaganda
5	0	0	507	1	0	d = recreation	3	1	0	194	0	1	d = recreation
0	1	0	0	76	0	e = regulation	2	0	0	0	78	0	e = regulation
2	0	0	0	0	20	f = reporting	14	0	0	0	0	29	f = reporting
Crisp BNC corpus (accuracy: 97%)							Fuzzy BNC corpus (accuracy: 86%)						
a	b	c	d	e	f	← classified as	a	b	c	d	e	f	← classified as
721	2	89	66	32	55	a = discussion	721	2	89	66	32	55	a = discussion
41	17	14	4	12	2	b = instruction	41	17	14	4	12	2	b = instruction
176	8	394	3	33	13	c = propaganda	176	8	394	3	33	13	c = propaganda
51	2	2	890	0	4	d = recreation	51	2	2	890	0	4	d = recreation
55	12	45	0	339	19	e = regulation	55	12	45	0	339	19	e = regulation
101	3	23	18	23	183	f = reporting	101	3	23	18	23	183	f = reporting
Russian fuzzy training corpus (accuracy: 74%)													

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 selected for 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 [21]. 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, as well as on the frequency of punctuation marks, e.g., quotes, exclamation and question marks each contributed to a feature. As shown in [22] this combination of parameters is known to be the most reliable indication of genres. 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. [23,24]. However, the chosen set does discriminate between the text types, some information (such as HTML tags) has been lost in the process of corpus creation, and, more importantly, this chosen combination of POS trigrams with punctuation marks is applicable to both traditional written texts and webpages.

Table 1 compares the result of training using a ‘crisp’ corpus against a ‘fuzzy’ corpus. 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 categories to examples outside clear-cut categories, when the classifier is applied to the rest

of the BNC. 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. The results reported below are based on fuzzy training corpora.

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 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, animacy, 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., *would* vs. *must*), in POS trigrams they are represented by a single tag.

Finally, the jungle of the Internet 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 the accuracy of classification can be attributed to three factors:³

- 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

³ The BNC has been retagged with TreeTagger, the same tool used for tagging the Internet corpus, so the tagset was exactly the same.

Table 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%

boilerplate, navigation frames, ASCII art. In spite the best efforts to remove this noise, the accuracy of automatic cleaning is below 75% [20].

- the language of the training corpus is to some extent different from the language used in traditional corpora, e.g., not only British English is included in the Internet sample, FAQs are organised differently from tutorials listed in the BNC, the core of BNC texts stems from 1980s (the accuracy on the Russian Internet sample was higher because the RNC is based on modern text and I-RU is much more homogeneous in terms of the dialects it contains).

4 Analysis of results

The results of the automatic assessment of the composition of traditional and Internet corpora are presented in Table 2. The composition of the 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 combat 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 the Internet corpora used, less for ukWac).

4.1 Qualitative assessment of texts in each category

Discussion This is the biggest category with a variety of subtypes according to the audience or publication medium. 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 are written for the professional audience as well as for the general public). This category also incorporates texts classified as ‘misc’, those that do not find their own category.

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. [3, 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. This genre is much more common in ukWac, as it was created by crawling webpages without any restrictions on their lexicon. Texts of this type are also important because of their potential to mislead POS taggers or other NLP tools. They often contain incomplete sentences, while the visual boundary between chunks (corresponding to sentences and obvious to human readers) is often lost in the process of creating a plain text version for storing in a corpus.

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_1f_e.asp

Texts that are trying to teach us something constitute about one quarter of the entire Web in English. This is the second most frequent text type (this finding is consistent for I-EN and ukWac). 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 Internet sample) is the greater difficulty of detecting them in Russian. According to the Russian confusion matrix in Table 1 the majority of texts classified as ‘instruction’ in the training set got reclassified as ‘discussion’. More attention is needed to finding features that can detect this class in Russian reliably.

Propaganda The common perception is that the Internet consists of porn and spam. For the coverage of different domains, including porn, see [25]. As for genres, spam is not very common on the Web (spam messages are too ephemeral to be recorded on webpages). The amount of texts with propaganda of various sorts is in the range of 12 to 16%, more in ukWac, much less in the BNC. Pages classified in this way typically promote goods and services, e.g., <http://www.dr911.com/>, which is not strictly speaking spam.

Recreation It is known from other studies [3] that texts written with the purpose of recreation, such as fiction, are rare on the English Internet (because of copyright restrictions), while they are quite frequent for Russian. The present

experiment confirms this view to a certain extent. However, such texts do exist in the two English Internet corpora. The classifier is actually quite generous in assigning this category to a text, e.g., http://42.blogs.warnock.me.uk/2006/05/cycling_fame.html, that describes an event and is written in a chatty style (descriptive texts are normally classified as ‘reporting’ otherwise). At the same time, one can argue that texts of this sort are reasonable to classify as aimed at recreational reading. At the same time, even for English, fiction proper is not entirely missing. 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.

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 1 their detection in English is easy for the SVM classifier, so the figure for English in Table 2 can be assumed to be reliable. As for Russian, there was only one text of this type in the random sample, hence the classifier cannot be trained reliably. As a result there are numerous Russian texts classified as ‘discussion’ that can be reasonably treated as regulatory documents, e.g., <http://www.dmpmos.ru/law.asp?id=30020>.

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, descriptions of historic events (wars or technological developments), etc.

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 3 lists genres automatically assigned to documents collected from one website devoted to a large international conference. The classification in all cases seems 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 the classification results. The first three newsletters promoted the conference or its sponsors, while the last one mostly consisted of an informative interview. The CVs in question were written in two different styles. One style describes

Table 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

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.

5 Conclusions and future research

This paper reports only the most preliminary study, which was aimed at uncovering the composition of the Internet jungle. Both the typology for webgenres and the set of features used to classify webpages automatically are 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 huge corpus like I-EN (160 MW) or ukWac (2 GW). In short the proposed procedure looks like this:

1. take a corpus which composition is known (source corpus);
2. train a classifier on its subset;
3. apply it to a sample of the 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. As an experiment, I classified I-EN using the 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]), though this was done without correcting the classifiers on a sample from I-EN. However, the value of such tests is limited, as the experiments with the BNC and RNC (Section 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 2 to say that the composition of the Internet is as follows: instructions – one quarter, propaganda – 10-15%, lists and catalogues – 5%, regulations – about 3%, etc. However, there are obvious limitations of using training corpora of 250 webpages. For instance, the Russian training sample contained just single examples 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 [3]), 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 likely to be classified as any other macrogenre. 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 Marina Santini’s PhD thesis [13, Ch. 11].

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,⁴ frequency bands, n-grams, lexical density, etc), grammatical features other than POS trigrams (especially since the latter are quite sparse in Russian), text statistics (average document or sentence length, web-specific markup statistics or URL components, etc), as well as into methods for more efficient population of the feature set with features corresponding to individual classes.

⁴ The use of keywords for genre detection has been studied, e.g., in [26] or [27].

The tools for genre classification described in this paper and the results of classifications of Internet corpora into genres are available from <http://corpus.leeds.ac.uk/serge/webgenres/>

References

1. Lee, D.: Genres, registers, text types, domains, and styles: clarifying the concepts and navigating a path through the BNC jungle. *Language Learning and Technology* **5**(3) (2001) 37–72
2. Kilgarriff, A.: The web as corpus. In: *Proc. of Corpus Linguistics 2001*, Lancaster (2001)
3. Sharoff, S.: Creating general-purpose corpora using automated search engine queries. In Baroni, M., Bernardini, S., eds.: *WaCky! Working papers on the Web as Corpus*. Gedit, Bologna (2006) <http://wackybook.sslmit.unibo.it>.
4. Joho, H., Sanderson, M.: The SPIRIT collection: an overview of a large web collection. *SIGIR Forum* **38**(2) (2004) 57–61
5. Baroni, M., Kilgarriff, A.: Large linguistically-processed Web corpora for multiple languages. In: *Companion Volume to Proc. of the European Association of Computational Linguistics*, Trento (2006) 87–90
6. Baroni, M., Bernardini, S.: Bootcat: Bootstrapping corpora and terms from the web. In: *Proc. of LREC2004*, Lisbon (2004)
7. Macdonald, C., Ounis, I.: The TREC blogs06 collection: Creating and analysing a blog test collection. Technical Report TR-2006-224, Department of Computing Science, University of Glasgow (2006)
8. Mehler, A., Gleim, R.: The net for the graphs - towards webgenre representation for corpus linguistic studies. In Baroni, M., Bernardini, S., eds.: *WaCky! Working papers on the Web as Corpus*. Gedit, Bologna (2006)
9. Halliday, M.A.K.: *An Introduction to Functional Grammar*. Edward Arnold, London (1985)
10. Sharoff, S.: Methods and tools for development of the Russian Reference Corpus. In Archer, D., Wilson, A., Rayson, P., eds.: *Corpus Linguistics Around the World*. Rodopi, Amsterdam (2005) 167–180
11. Ferraresi, A.: Building a very large corpus of english obtained by web crawling: ukwac. Master's thesis, University of Bologna (2007)
12. Biber, D.: *Variations Across Speech and Writing*. Cambridge University Press (1988)
13. Santini, M.: Automatic Identification of Genre in Web Pages. PhD thesis, University of Brighton (2007)
14. Martin, J.R.: Language, register and genre. In Christie, F., ed.: *Children Writing: reader*. (ECT Language Studies: children writing). Deakin University Press, Geelong, Vic. (1984) 21–30
15. Kessler, B., Nunberg, G., Schütze, H.: Automatic detection of text genre. In: *Proceedings of the 35th ACL/8th EACL*. (1997) 32–38
16. Braslavski, P.: Document style recognition using shallow statistical analysis. In: *ESSLLI 2004 Workshop on Combining shallow and deep processing for NLP*. (2004) 1–9
17. Jakobson, R.: Linguistics and poetics. In Sebeok, T.A., ed.: *Style in Language*. The M.I.T. Press (1960) 350–377

18. EAGLES: Preliminary recommendations on text typology. Expert Advisory Group on Language Engineering Standards document EAG-TCWG-TTYP/P (1996)
19. Sinclair, J.: Corpora for lexicography. In Sterkenberg, P.v., ed.: *A Practical Guide to Lexicography*. Benjamins, Amsterdam (2003) 167–178
20. Baroni, M., Chantree, F., Kilgarriff, A., Sharoff, S.: Cleaneval: a competition for cleaning web pages. In: *Submission to Proc. of the Sixth Language Resources and Evaluation Conference, LREC 2008, Marrakech (2008)*
21. Witten, I., Frank, E.: *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco (2005)
22. Santini, M.: Linguistic facets for genre and text type identification: A description of linguistically-motivated features. Technical Report ITRI-05-02, University of Brighton (2005)
23. Allen, P., Bateman, J.A., Delin, J.L.: Genre and layout in multimodal documents: towards an empirical account. In Power, R., Scott, D., eds.: *Proceedings of the AAAI Fall Symposium on Using Layout for the Generation, Understanding, or Retrieval of Documents*, Cape Cod, Massachusetts, American Association for Artificial Intelligence (1999) 27–34
24. Meyer zu Eissen, S., Stein, B.: Genre classification of web pages. In: *Proceedings of the 27th German Conference on Artificial Intelligence, Ulm, Germany (2004)*
25. Sharoff, S.: Classifying web corpora into domain and genre using automatic feature identification. In: *Proc. of Web as Corpus Workshop, Louvain-la-Neuve (September 2007)*
26. Xiao, Z., McEnery, A.: Three genres in modern American English. *Journal of English Linguistics* **33**(1) (2005) 62–82
27. Crossley, S.A., Lowerse, M.: Multi-dimensional register classification using bigrams. *International Journal of Corpus Linguistics* **12**(4) (2007) 453–478