

In the Garden and in the Jungle: Comparing Genres in the BNC and Internet

Serge Sharoff¹

¹University of Leeds (UK)

Short Abstract

According to Adam Kilgarriff the BNC is a jungle when compared to smaller Brown-type corpora, but it looks more like an English garden when compared to the Internet (Kilgarriff and Grefenstette, 2003). In this presentation I will compare English and Russian Internet corpora against their human-collected counterparts using two methods: the first involves manual annotation of a subset of Internet corpora, the second one uses probabilistic classifiers. The study shows that the Internet is not radically different from the BNC: Internet corpora do contain a wide range of genres and approximate many genres that exist in their printed form, the same is true for the audience level (texts for professional or layman texts).

Long Abstract

Unlike traditional representative corpora (e.g. the BNC), large corpora automatically collected from the Web (Joho and Sanderson, 2004, Sharoff, 2006) lack important information documenting them, such as their domains and genres. The task of classifying their texts and comparing their composition to traditional corpora is difficult for several reasons. First, no established classification of genres exists: 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. Third, we need reliable automatic methods for identifying genres of arbitrary webpages. The fourth problem concerns the very design of the genre inventory. If the goal is to classify every Webtext, 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 conventions applicable to the humanities, engineering or fundamental research in the 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, for instance, blogs, personal homepages or spam. To compare them to traditional sources we need a common denominator, such as communicative intentions underlying creation of texts in respective genres.

A starting point for establishing a set of communicative intentions can be taken from such studies as (Sinclair, 2003, Sharoff, 2004, Aires et al., 2005):

1. **recreation** – such texts are written for leisure-time reading; the two important subclasses are fiction (science fiction, crime, etc) and nonfiction (biographies, memoirs, etc);

2. **information** – such texts provide information about something and answer questions on what has happened, and how or why it happens; newswires and encyclopedic entries are typical examples of this text type;
3. **instruction** – such texts explain how to do something; e.g. recipes, software man pages, etc, as well as more descriptive texts such as FAQs, tutorials and textbooks;
4. **discussion** – such texts are aimed at discussing a state of affairs (including typical newspaper articles, academic papers, travel stories, etc); unlike purely informative texts they tend to present the opinions of their authors;
5. **recommendation** – the purpose of such texts is to make you behave in a certain way; examples include propaganda and advertising.

In addition to this a study of genres should use parameters other than communicative intentions, e.g. the authorship (single or corporate), audience, publication medium, etc. Experiments with users show that detection of these more abstract parameters is considerably more difficult than detection of simple genre labels by their look and feel. For instance, users know how a blog looks like, so if a page looks like a blog, it can be classified as a blog, whereas the choice between information and discussion, as the main aim of its creation is less obvious. However, if a list of genres includes a simple entry for blogs, it cannot be compared to anything in the BNC, whereas their function is similar to that of opinion columns in newspapers, and is different from them in the audience size, distribution mode and authorship. Also a list of news items with the possibility of leaving reader's comments looks very similar to blogs in its layout, but does not share the communicative function with them. This suggests the need to ensure the right balance between abstract parameters and look-and-feel features.

The second step of this study was to compare the distribution of genres in the BNC and RNC against their Internet counterparts. For this we mapped the genre labels used in the BNC and RNC to the more general categories listed above, for instance, academic and non-academic papers can be treated as 'discussions', fiction, biographies and popular lore as 'recreational' texts. Not all genre labels can be mapped unambiguously, so this gives us 828 BNC texts to work with (3,648 texts were used for Russian, as RNC texts are considerably shorter). Then, we trained SVM classifiers on the frequency of POS trigrams describing individual texts, as well as the frequency of punctuation marks. As shown in (Santini, 2005) this is known to be the most reliable indication of genres, which is applicable to both traditional written texts and webpages. For English the procedure achieved 94% accuracy with 10-fold cross-validation, while being significantly lower (76%) for Russian, which can be explained by the free word order which makes POS trigram statistics sparser, especially on shorter texts. Finally, we applied the models trained on the BNC and RNC to English and Russian texts from Internet corpora. The accuracy was tested on smaller samples of 200 webpages annotated manually, see (Sharoff, 2006). The accuracy drops (to 82% for English, 59% for Russian), but this still allows reasonable interpretation.

The results of the experiment show that recreational texts are seriously under-represented in Internet corpora: 27% in the BNC, 43% in the RNC, but only 4% in I-EN and 11% in I-RU (because of the larger number of pirated texts and exchanges of jokes in the Russian Internet). At the same time, the balance of other text types in the BNC and RNC is reflected in Internet corpora. The implication of these results for constructing a reference corpus of web genres is

that they indicate the relative proportion of text types in the Internet and guide towards figures for balancing the reference corpus.

References

- Aires, R., Santos, D., and Alusio, S. (2005). "Yes, user! ": compiling a corpus according to what the user wants. In *Proc. Corpus Linguistics*.
- Joho, H. and Sanderson, M. (2004). The SPIRIT collection: an overview of a large web collection. *SIGIR Forum*, 38(2):57–61.
- Kilgarriff, A. and Grefenstette, G. (2003). Introduction to the special issue of the web as corpus. *Computational Linguistics*, 29(3):333–347.
- Santini, M. (2005). Linguistic facets for genre and text type identification: A description of linguistically-motivated features. Technical Report ITRI-05-02, University of Brighton.
- Sharoff, S. (2004). Towards basic categories for describing properties of texts in a corpus. In *Proceedings of the Forth Language Resources and Evaluation Conference, LREC 2004*, Lisbon.
- Sharoff, S. (2006). Creating general-purpose corpora using automated search engine queries. In Baroni, M. and Bernardini, S., editors, *WaCky! Working papers on the Web as Corpus*. Gedit, Bologna. <http://wackybook.sslmit.unibo.it>.
- Sinclair, J. (2003). Corpora for lexicography. In Sterkenberg, P. v., editor, *A Practical Guide to Lexicography*, pages 167–178. Benjamins, Amsterdam.