Open-source Corpora

Using the net to fish for linguistic data

Serge Sharoff
University of Leeds

The paper proposes a methodology for collecting "open-source" corpora, i.e. corpora that are automatically collected from the Internet and distributed in the form of a list of links with open-source software for recreating their full text. The result is a random snapshot of Internet pages which contain stretches of connected text in a given language. The paper discusses a methodology for acquiring such corpora, two ways of documenting them (using a set of metatextual categories and by comparison to frequency lists from existing corpora) and their function as benchmarks for comparing results of linguistic inquiry. Experiments with a variety of languages show that Internet-derived corpora can be successfully used in the absence of large representative corpora that are rare and expensive to build.

Keywords: Internet, corpus composition, representative corpora, frequency lists

1. The problem

Lexicographic studies using corpora can be reliable only if corpora providing the basis for the study are sufficiently large and diverse. The famous example with collocations of *powerful* and *strong*, such as *strong tea* (Halliday 1966:150), can only be studied computationally on a corpus of at least the size of the *British National Corpus* (BNC). In 100 million words of the BNC, the expression *strong tea* occurs 28 times,[1] which makes it a reasonably strong collocation along with *strong* {*candidate, contrast, leadership, reason*}, all of which have roughly the same frequency and statistical significance according to the log-likelihood score. However, the chances of detecting these collocations in a smaller corpus are minuscule: *strong tea* occurs only once in the *Brown* corpus, and it contains no instances of *strong candidate, leadership* or *reason*.

International Journal of Corpus Linguistics 11:4 (2006), **435–462**.
ISSN 1384–6655 / E-ISSN 1569–9811

The wide availability of the BNC is one of the reasons why corpus-based research is done mostly for English, as many academic researchers do not have access to sufficiently large corpora in other languages, especially for lesser-studied languages such as Romanian or even Chinese and Russian. Existing studies in such languages are typically based on small corpora, which limit the possibility of lexicographic research for many moderately frequent lexical items, or on opportunistic text collections, which do not have sociolinguistic validity.

The present study experimented with the development of Internet corpora for Chinese, English, German, Romanian, Russian and Ukrainian. There are no BNC-type corpora for Chinese and Romanian currently available. Existing text collections are either small or opportunistic. The situation with German looks slightly better: there are two German corpora comparable to the BNC in size. First, there is an opportunistic corpus from *Institut für Deutsche Sprache* (IDS), which is huge: the main written corpus contains almost 1 billion words. However, it is not representative, as it is clearly biased towards newspaper texts. For instance, the word *SPD* (the name of a German political party) is more frequent in it than *ja* (yes), *Kinder* (children) or *Frau* (woman). IDS also collected DEREKO (*Deutsches Referenzkorpus*), which is probably better balanced, but it is not available for research. Another corpus was collected within the project *Das Digitale Wörterbuch der deutschen Sprache* (DWDS). The DWDS corpus is more balanced, but very little information is available about its text collection (unlike the data about the composition of the BNC). For Russian there is an ongoing project for collecting the *Russian Reference Corpus* (RRC), a representative corpus of 100 million words (Sharoff 2004). Its preliminary results (a corpus of 35 million words, almost half of which belong to the domain of fiction) are used in this study for comparative purposes. However, none of the above mentioned corpora is available as files in the same way as the BNC is available for English, which you can process using your own tools, for instance, to study statistical properties of texts, sentences or collocations, to produce frequency lists, vector lists for a word, etc.

There are claims that large corpora are increasingly available for major languages. For instance, *Reuters* released their corpus of newswires from 1996–1997 amounting to about 100 million words. There are also "Gigaword" newswire corpora for Arabic, English and Chinese.[2] Large newswire corpora can be useful for many purposes, such as detection of named entities, compiling lists of terms, testing of information retrieval algorithms, etc. However, they are representative for only one type of language: the language of newswires taken from one or two news sources. This restricts the range of syntactic

constructions and uses of words to formal description of past events, cf. a comparison of the *Reuters* lexicon to that of the BNC in Section 3.2 below.

A less opportunistic approach as exemplified by the BNC assumes a choice of targets for the corpus composition and selection of texts to meet those targets. This ensures that a limited sample of texts stored in the corpus is representative for the language as a whole. However, collecting BNC-like corpora is quite an expensive enterprise, which is hard to repeat for another language or for the same language to take into account language change. The BNC is based on texts from the 1970s, 1980s and early 1990s, so it does not show many examples of contemporary language use. For instance, the BNC does not contain any use of the word *browser* in the sense of an Internet browser. Development of BNC-like corpora also encounters the problem of the freedom to distribute the resulting resource: publishers are less and less inclined to give the right to distribute the corpus in the source form. On the other hand, publicly available representative corpora like the BNC are important as a benchmark: they provide a basis for comparing results obtained by independent researchers. As a randomly chosen example, Szmrecsanyi (2003) relates his study of expressions of futurity in English to a study of g*onna* and *going to* by Berglund (2000) (both studies are, to a large extent, based on the BNC). The BNC also provided the basis for describing the composition of Internet corpora in this study.

The use of Web as a source of linguistic data has been a championed by Kilgarriff (2001). However, his idea of creating a Linguistic Search Engine has not been put into practice so far. The aim of this paper is to pursue the same line of research and evaluate its results. In the following section, we propose a methodology for collecting BNC-like representative monolingual corpora by making a random snapshot of the current state of the Internet in a given language. When we refer to an "Internet corpus" throughout the paper, we take this to be a corpus collected according to the proposed methodology. Then we discuss the composition of Internet corpora and estimate their balance by comparing them to existing resources for these languages. We employ two methods:

1. annotation of their samples using a text typology and comparison of the results to other annotated resources (such as the BNC)
2. building frequency lists and comparing them with frequency lists taken from other corpora.

After surveying related research, we finally proceed to conclusions concerning the quality of Internet-derived corpora and the perspective for using them in linguistic studies and language teaching, in particular, we discuss the possibility of collecting "open-source" reference corpora that are distributed in the form of URL lists.[3]

2. Corpus collection methodology

According to the proposed methodology a large reference corpus for an arbitrary language can be collected in six steps: word selection, query generation, downloading from the Internet, post-processing, composition assessment and comparison of word lists (the last two steps are optional).

2.1 Step 1: Word selection

In this step we select 400–500 words from the list of the most frequent word forms in a given language. There is no need to select words that are unique to that language, as the next step uses the language filter of a search engine. To be considered as "good", query words must be sufficiently general, i.e. they do not belong to a very specific domain. This ensures that randomly generated queries do not highlight a particular topic. For instance, if a word like *Tory* was used in the word list, the resulted corpus would be biased towards British political news and discussions. Words *events* and *picture* can be included in the list for collecting a corpus for English, because the word *events* occurs on webpages referring to business meetings, races, performances, war reports, etc. Similarly, *picture* occurs mostly in constructions like *this isn't the whole picture; in the larger picture* and actually it brings very few pages specifically in the domain of arts. At the same time function words are general, but they do not make good query words, as they are frequently used in pages that do not contain complete sentences, such as catalogues or lists of headlines, which are not good candidates for a corpus. Also search engines often treat function words differently from common words by including them in stop lists or indexing them in a special way.

Since the Google interface (used in Step 2) does not perform lemmatisation, we have to rely on a list of word forms only. For languages with elaborate morphology, such as Arabic, Romanian or Russian, there can be 10–20 word forms per lemma (even more in Czech). Thus, a list based on exact word forms in such languages operates with query terms of lower frequency. For instance, two lemmas *event* and *событие* are good translation equivalents having roughly the same rank and frequency in English and Russian, as their position in the frequency list is around 500, and their frequency is about 200 instances per million words (ipm). However, the frequencies of the exact forms *event* and *событие* are quite different: 104 ipm for *event* vs. 22 ipm for *событие*. However, this difference did not create a problem for our study, because many webpages exist in those languages anyway, so we can find sufficient number of

hits for each query. In the end, Internet corpora we collected for languages with rich morphology (Romanian and Russian) exhibit frequency patterns similar to languages that have fewer word forms per lemma (see the results in Section 3.2). Čermák and Křen (2005) in their study of the frequency distribution in English, Czech and Russian corpora also report that the general principle of the Zipfian distribution is common to these three languages, even if its exact parameters are language-specific.

In our experiments we used frequency lists taken from existing corpora, such as the BNC or RRC. If we want to develop a corpus for a language for which no corpora are currently available at all, we can rely on intuition in creating the word list for queries, because the exact frequency of words is less important than selection of common frequent words that do not point to a specific domain.

2.2 Step 2: Query generation

In this step we produce a list of queries each of which consists of four words chosen at random from the list selected in Step 1, send the queries to a search engine (such as Google or Yahoo) and store the top URLs returned by the search engine. Search engines can restrict the search to a variety of languages using their own linguistic filters. If the language for which we want to collect a corpus is not covered, each query can be complemented with a couple of very frequent function words that are not used in cognate languages, e.g. for Ukrainian we used the query *має* OR *її* (has OR her).

The condition for using four common words in a query follows the requirement to get pages that contain relatively long pieces of connected text, with a smaller number of "noisy pages" in the form of price lists, tables, lists of links, etc. The presence of one-two common words in a query does not guarantee an instance of connected text. For example, the first page retuned by Google for the query *picture* AND *extent* is a photo with a short advertisement containing the word *extent*. At the same time, a four-word query is much more likely to yield a page useful for a corpus. For instance, the top ten pages produced by the query *picture* AND *extent* AND *raised* AND *events* all have stretches of narrative prose ranging from two to five thousand words (not counting navigation frames). The pages retrieved by this query also refer to a variety of domains, including pages on war crimes, public affairs education, astronomy, medicine, return-on-investment study, etc.

However, if we use queries longer than four words, the number of pages returned is much smaller, so that the result will not qualify as a random

snapshot of the Internet. Even for English (the language most widely used on the Internet) a query of eight words frequently produces no result at all or the result consists of duplicate pages. It is possible to relax the condition for four words in a query for languages which do not have sufficient number of Internet pages. For instance, we used queries of three words for collecting the Romanian corpus. The reason for this was that we wanted to collect a corpus with proper encodings of diacritics, but they are not frequently available on Romanian Internet pages.

2.3 Step 3: Downloading

For each query Google returns 10 URLs, which are used for further processing. In the current setup we used 5,000 queries, which resulted in 50,000 URLs. However, some URLs can be picked up more than once as a result of different queries. The downloading step reduces the number of URLs further, because of the dynamic nature of the Internet: not all pages indexed by Google are available at the time of downloading. This may require additional queries to extend the database of URLs. The target of collecting 40,000–50,000 URLs has been chosen after several experiments as a threshold for producing a reasonably large corpus of more than 100 million words of real narrative texts. This makes the size of the corpus comparable to the BNC. The procedure can be repeated to enlarge the corpus up to the limit of all texts in this language indexed by the search engine, but for lexicographic research a corpus of 100 million words gives sufficient evidence for the top 25,000 words (which have at least 100 occurrences in this case), while the upper limit depends on what is a reasonable size for corpus storage and reasonable time for producing concordances. The list of successfully downloaded URLs is stored in the corpus database and can be used to recreate the corpus by other researchers.

2.4 Step 4: Post-processing

Pages collected in the previous step are subjected to automatic postprocessing. First, it is necessary to unify the page encoding, which is also not always specified in the page attributes (Russian pages can come in 6 different encodings for Cyrillic characters). Second, we convert pages from HTML into plain text and remove navigation frames (following Finn et al's (2002) idea about the detection of extensive use of links). Finally, we filter out pages that are either completely identical (e.g. two copies of the GNU Public License) or almost identical (e.g. a page with navigation and its printer-friendly version). This results in a

clean corpus in plain text format. In order to make it a proper corpus, we need language-dependent morphosyntactic processing, such as tokenisation (more important for Chinese and other Asian languages), lemmatisation (especially for Slavonic languages), as well as part-of-speech tagging and further syntactic and semantic processing, if respective tools are available.

2.5 Step 5: Composition assessment

The usability of a corpus collected automatically depends on our understanding of what the corpus is composed of. We can assess the composition of a new corpus by developing a text typology and annotating texts in the corpus according to this typology. However, the procedure for assessing the composition of a large corpus has to balance between what is theoretically sound and what is practical. As it is not reasonable to assess manually 40,000–50,000 texts, it is necessary to choose a representative sample: a subcorpus of N texts which can be coded in reasonable time and provide statistically sound results about the composition of the whole corpus.

Textbooks on statistics offer a straightforward procedure which estimates how well a random sample represents the total population (Upton & Cook 2001:301). It uses two parameters:

- σ, the symmetric interval desired for the accuracy of sampling, i.e. we want our results to be valid with the precision of ±σ;
- *p*, the confidence that the results that could be obtained from the total population are indeed within the symmetric interval obtained from studying the sample (even if sampling is random, there is always a chance that the properties of the sample are different from those of the rest of the corpus).

According to statistical tables, to achieve the symmetric intervalσ=5% with 90% confidence we need a sample of about 200 documents. For instance, if we studied a sample of 200 documents and the number of texts written by men in our sample is 30%, then we can claim with 90% confidence that the number of such texts in the total corpus is in the range of 25–35%. A better approximation within the interval of ±1% with 95% confidence will require a much larger sample, of about 1,500 documents.

In our experiments we used samples consisting of 200 documents for each corpus. A sample of this size is reasonable to assess according to a number of text description categories. However, as the accuracy of estimation for each category is ±5%, categories that occur in less than 5% of cases in the sample cannot be assessed reliably. For coding we used the Systemic Coder (O'Donnell

1995), which prompts for values of categories and allows basic statistical analysis of the results.

2.6 Comparison of word lists

Assessment of the corpus composition involves a significant amount of manual coding and implies near-native knowledge of the language and culture for which the corpus has been created. If the language for which we created the Internet corpus has any benchmark corpus with known composition, such as a reference or newswire corpus, we can assess the major differences between the newly acquired corpus and the known benchmark corpus by means of studying frequency lists. Since the lexicon of a corpus reflects its content, this will give us an aggregate assessment of differences between the two corpora (provided that the statistical measure we use to compute the significance of the frequency difference is theoretically sound). The analysis of differences can suggest ways in which one corpus is less balanced than the other. For instance, if *cost, bank* and *company* are significantly more frequent in one corpus in comparison to the other one, then we can assume that the first corpus contains more financial news.

In our study we calculated log-likelihood values for the frequencies of words in a pair of corpora, took words with the highest values and listed separately words that are more frequent (overused) and less frequent (underused) in the second corpus in comparison to the first one. For more information on the computational aspects of the procedure, see Rayson and Garside (2000).

3. Corpus evaluation

3.1 Evaluation of corpus composition

Assessment of the corpus composition requires a text typology to annotate texts in the sample. Existing research within corpus studies has produced two theoretically sound classifications. First, the Text Encoding Initiative (TEI) provides a very extensive set of tags and attributes for encoding text headers. However, the full TEI set is huge and many TEI tags are irrelevant for the purposes of corpus development. For instance, some of them, such as editor, publisher or distributor, are aimed at documenting texts for library records. At the same time the TEI guidelines are not specific enough, because they lack a text typology proper, such as taxonomies of basic problem domains or properties

of the intended audience. A TEI-based text typology was developed to encode files in the BNC, but it paid more attention to the bibliographic classification of corpus files and did not touch some issues concerning the function a text carries in the linguistic community. The BNC typology also mixes some natural dimensions of classifications. For instance, spoken texts in the BNC do not have domain categories, e.g. life, art or politics, even though the latter are legitimate topics of conversation.

Second, the European Advisory Group on Language Engineering Standards (EAGLES) produced text typology guidelines in work headed by John Sinclair (EAGLES 1996; Sinclair 2003). However, EAGLES guidelines (unlike TEI) do not define a set of tags and attributes. More importantly they do not always deal with text types that are frequent in general-purpose corpora or webpages, such as types of newspaper texts or advertisements. Finally, TEI guidelines, and the text typologies from the BNC and EAGLES offer too many options in the sense that if we use all the categories available for coding even a sample of a corpus, the coding will take a lot of time. Furthermore, the values for many options are not known when coding webpages.

We attempted to develop a small set of categories and rules for assigning values of those categories. This set of proposed categories is specific enough to describe the great majority of Internet pages with adequate sociolinguistic precision, but at the same it is quite small to allow rapid sampling of a statistically significant corpus chunk (of about 200 documents).

Another requirement for the set of categories is the reliability in detection of their values on the basis of information provided in Internet pages. For instance, the sex of the author can be reliably identified in the languages used in the study by his/her first name, if it is given, e.g. *John* vs. *Mary*. There are relatively few cases, when it cannot be done, either because it is ambiguous, like *Chris* in English, or the sex association is not known to the coder, like *Cody*. The sex of an unknown author sometimes can be guessed from semantic clues, e.g. if the author refers to *my husband,* or from grammatical properties, such as gender agreement in Russian (*я была…* — I was-fem). At the same time detection of the age of the author or the size of the audience is much less reliable, so they were left out of our classification scheme.

We assess each text in the sample in terms of its authorship, mode (or channel), knowledge expected from the audience, the aim of text production and the generalised domain. In addition to the set of categories we provide explicit instructions for filling their values on the basis of observable features of texts.

The results of assessment of the composition of automatically acquired corpora are shown in Table 1 in the Appendix (Internet corpora are abbreviated

there as I-EN, I-RU and I-DE for English, Russian and German, respectively). The English and Russian Internet corpora can be also compared against data obtained from representative corpora for those languages, though the comparison cannot be complete, as neither BNC nor RRC classify pages with respect to the purpose of their production.

3.1.1 *Authorship*

Information about the authorship uses the following values:

- **single** — created by a single named author, we also classify the sex of explicitly named single authors, in so far as this can be detected using the name and other lexical or syntactic clues;
- **multiple** — created by several named co-authors;
- **corporate** — created by a corporate author (in this case there is a corporate copyright statement and a human author is not given);
- **unknown** — no information about the author is available on the page and can be inferred without significant extra efforts.

In comparison to traditional representative corpora, Internet corpora contain significantly more texts coming from corporate sources (44% for I-EN vs. 18% for the BNC), while they consistently underrepresent female writers (23% of texts in I-EN are written by men vs. just 3% by women in comparison to the 28% vs. 13% split in favour of male writers in the BNC).

3.1.2 *Mode*

The classification of texts with respect to their mode follows the EAGLES guidelines using the following values:

- **written** — traditional written texts, including newspapers, homepages, etc;
- **spoken** — transcripts of sound-wave recordings;
- **electronic** — spontaneous communication, such as emails, electronic forums or chat rooms.

The EAGLES guidelines introduced the 'electronic mode' "to emphasise that language transmitted in electronic media is not quite the same as the older established modes". For the purposes of coding webpages (all of which exist in electronic form), the use of the 'electronic mode' was restricted to spontaneous electronic communication. The separation is important because in comparison to traditional written texts they are similar to spoken communication in the spontaneity of their production (like face-to-face or telephone conversations).

However, they are *not* spoken texts, so they lack prosodic information, which is compensated by capitalisation or new means of expression, such as emoticons and smileys. Electronic texts also exhibit a large number of typos and grammatical errors.

Only 10% of the BNC consists of spoken texts, because collection of a large spoken corpus was not considered to be practical. In Internet corpora we find very few cases of transcripts of spoken language, but spontaneous language is predominantly represented by discussion forums, so electronic texts correspond to 16% of the Internet corpus for Russian, 13% for English and 9% for German.

3.1.3 *Audience*

It is frequently impossible to make a reliable judgement with respect to values of the audience parameters using the full set of categories from the BNC and EAGLES text typologies. For instance, the BNC index uses identical codes for describing an article on assimilation and adaptation in mental health care from *The British Journal of Social Work* (text GWJ) and an article on French smoking habits from the tabloid *Today* (CEK): both are published in periodicals and belong to the domain of humanities. The BNC typology provides a code distinguishing the audience level, but both texts are coded as medium.

In our experience the judgement on such audience parameters as its size or level are hard to make, but we can reliably code the level of *knowledge* expected from the audience to read a text:

- **general**, e.g. a text on ulcers from the BBC website; such texts are written in a way that anyone can read it if interested, they refrain from using terminology that the general public is not expected to know;
- **informed**, e.g. a description of ulcers for medical students; such texts are not very technical, but they do use a significant amount of specialist terminology;
- **professional**, e.g. an article in the Journal of Gastroenterology and Hepatology.

The exact boundaries between texts aimed at the general, informed or professional audiences are vague, but in the vast majority of cases the decision is clear. Internet corpora contain a good balance between these three categories, with the prevalence of texts being aimed at informed audiences, e.g. 33% for general, 45% for informed, 22% for professional texts in I-EN.

3.1.4 *Aims of text production*

We use a modified set of classes from Sinclair (2003) to code the aim of text production as:

- **discussion** — texts aimed at discussing a state of affairs (e.g. background articles in newspapers, academic papers, travel stories);
- **information** — Sinclair (2003) restricts the category to reference compendia, while in corpora we find such cases as: **reference** (dictionaries, encyclopedias), **data** (police reports, summaries, minutes of project meetings, etc), **news-reports** (e.g. a message informing about an earthquake differs from a newspaper article about rescue efforts, the latter being classified as **discussion**);
- **recommendation** — recommendations differ from discussions as they provide an incentive for doing or abstaining from doing something; examples of subclasses are: **advice**, **legal**, **advertisement**;
- **recreation** — the primary purpose of writing such a text is for leisure-time reading; the two important subclasses are **fiction** and **nonfiction**, further subclasses of fiction and nonfiction can be distinguished, but they are too rare on the Internet to warrant this;
- **instruction** — such texts are aimed at teaching their readers; the following subsclasses can be used: **manual** (e.g. recipes, flat-pack assembly or software man pages; they typically come in the form of itemised lists), **practical-how-to** (this category encodes more descriptive text varieties in comparison to manuals, the most frequent type of this category among Internet texts is an FAQ), **textbook** (on the Internet we typically have not complete textbooks, but explanations and introductory material on various topics, e.g. a Perl tutorial);

There are some vague cases between 'discussion' and 'recommendation', but in the majority of pages the distinction is clear: if it is evident that a text tries to persuade the reader as a potential customer or supporter, it is classified as 'recommendation', a text without obvious propaganda is 'discussion'.

A classification of this sort is used in neither BNC nor RRC, so Internet corpora have no basis for comparison. However, the three Internet corpora being compared are quite similar with respect to aims of their production. Internet texts most typically discuss a topic or give recommendations (most typically by advertising products and services).

Texts aimed at 'recreation' are treated as an important category in traditional corpora (fiction constitutes 17% of the BNC and 49% of the pilot version of the RRC, though the latter figure will be lower in the final version). However,

because of copyright restrictions fiction texts are relatively rare on the Internet (especially in English and German, where they constitute just 3–4% of respective corpora). Texts aimed at recreation are more frequent in I-RU (11%), including fiction texts and exchanges of jokes, but still they are relatively rare.

3.1.5 *Domain*

The EAGLES guidelines mention the frequent variation of topics within a single document or conversation and reject the applicability of any general classification system (such as Dewey Decimal Classification). Instead, they list domains considered in various terminology and corpus studies and refer to the unsuitability of "trying to arrange a hierarchy of simple topic labels" (EAGLES 1996). However, in practical terms their list of some 30 domains is too fine-grained. What is more, a webpage can be a subject to a more delicate classification, which, nevertheless, should start from a node in the hierarchy.

Even though any classification of topics is not complete, we propose to use eight general categories for classifying webpages:[4]

natsci (maths, biology, physics, chemistry, geo, …)
appsci (medicine, computing, ecology, engineering, military, transport, …)
socsci (law, history, philosophy, psychology, sociology, language, education, …)
politics
business
life (a general topic that is used for fiction, conversation, etc.)
arts (visual arts, literature, architecture, performing arts)
leisure (sports, travel, entertainment, fashion…)

The labels associated with categories whenever possible follow the practice of the domain codes used in the BNC, but some have been changed to reflect additional dimensions of classification (e.g. spoken texts can have a legitimate domain) or to generalise over topics (e.g. 'life' incorporates imaginative texts, chats on love affairs or a weblog that reports parenting of a child). There is a natural trend to add more top level categories. For instance, such topics as sports and travel correspond to separate sections in newspapers. Similarly, there is a difference between subdomains within applied sciences and the humanities, e.g. medicine or law can be considered as candidates for top level categories, and so on. However, there is a danger in the infinite extension of the set, which in the end can explode into an EAGLES-like list. So in this attempt we decided to keep the minimal possible set of coherent categories.

The comparison of Internet corpora to the BNC and RRC reveals underrepresentation of texts from arts and humanities ('arts' and 'socsci') in Internet

corpora in comparison to their traditional counterparts, e.g. 16% for 'socsci' in the RRC vs. 5% in I-RU. Even though the figures for English look closer (17% in the BNC vs. 16% in I-EN), the vast majority of texts considered as 'socsci' in the English Internet are legal texts (legislation, law reports, terms and conditions, etc), not texts in history, linguistics or education as in the BNC. At the same time there are many more texts from technical fields ('appsci') on the Internet: 7% in the BNC vs. 29% in I-EN (Internet texts most frequently belong to subdomains of computer science, medicine or construction industry). In comparison, 56% of the *Reuters* corpus consists of financial news (its C, E and M subcategories), while less than 0.5% is classified as science (GSCI), which includes 'natsci', 'appsci' and 'socsci' categories taken together. What is more, texts in the *Reuters* corpus are obviously not aimed at discussing scientific topics or teaching about them, but mostly give information in the form of news reports (Section 3.1.4 above).

3.2 Comparison of frequency lists

As assessment of corpus composition even using a limited sample is a time-consuming enterprise, we can describe a corpus by comparing its word list against word lists from existing corpora with a known composition.

Figure 1 shows triangulation of Internet corpora (corpora with unknown composition) with respect to two types of corpora: reference corpora (BNC and RRC) and newswire corpora (*Reuters* and *Xinhua*). Words that are less frequent in newswire corpora are shown in the left parts of Table 2 (see Appendix), words that are more frequent in newswire corpora are in the right parts. For the sake of space, the tables show only the 10–12 words with the most significant log-likelihood scores, but in examples we occasionally discuss some other words with high scores.

First we take two corpora with known composition and compare the frequency list of a newswire corpus (*Reuters*) against a representative corpus of general language (BNC). In this step we identify the differences between the lexicon of a representative corpus vs. the lexicon of a newswire corpus.

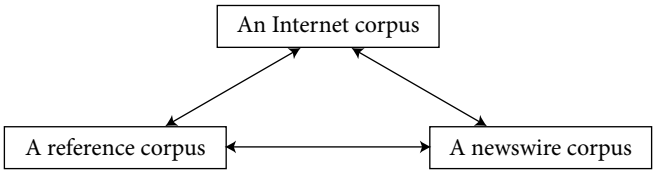

Figure 1. Triangulation for Internet corpora

Second, we compare an Internet corpus against a newswire corpus with known composition (English and Chinese Internet corpora against respectively the *Reuters* and *Xinhua* corpora). In this step we also compare the German Internet corpus against the IDS corpus, the composition of which is unknown, but it is likely that IDS exhibits some features of a newswire corpus. In doing this comparison we will try to show that Internet corpora differ from newswire corpora in more or less the same way as the BNC differs from the *Reuters* corpus.

In the third step, we compare two representative corpora with known composition (BNC and RRC for English and Russian) against their Internet counterparts to study the differences between the language use on the Internet and in general-purpose corpora. Word forms with the highest log-like scores are shown in Table 3 (see Appendix). Word forms were used instead of lemmas because of differences in the lemmatisation procedures used to produce frequency lists for the two reference corpora and the automatically acquired Internet corpora. This boosts differences in lemma lists significantly without any underlying linguistic reason.

Table 2 shows that newswire corpora in comparison to both the Internet and BNC overuse words referring to financial data (*million, Mark,* 经济), specific entities and institutions (*market, dpa,* 新华社), other financial terms (*share,* also *analyst, trader, price*) and exhibit greater use of temporal markers that specify the date and time of an event (*Tuesday, Uhr,* 日). Another specific feature of newswires is much greater use of reported speech, which is reflected in the overuse of such words as *say, sagen,* 表示. The latter word is actually restricted to the news reports register, while indirect speech in everyday life is typically reported using other verbs, such as 说 or 告诉. In German *sei/seien* (the subjunctive forms of *sein,* 'to be') are also markers of reported speech, in particular, they are frequently used as copular verbs in this context, for example:

> Jacques Delors pflegte zu sagen dass der Markt kurzsichtig sei und es deshalb politisch notwendig sei die Unterschiede zu verringern.
>
> [Jacques Delors was accustomed to saying that the market was short-sighted and hence it was politically necessary to reduce the disparities.]

The comparison of frequency lists confirms the intuition that the balance of the German IDS corpus is shifted towards the newswire use in comparison to the Internet corpus, as its frequency profile follows those of other corpora known to come from newswires (though the values of log-likehood scores suggest that newswire words are less prominent in the IDS corpus).

At the same time words that are *less* frequently used in newswire corpora in comparison to the BNC and Internet corpora include fewer first and second personal pronouns, question words (*what, welche,* 什么), modals (*can*, *muss*, 着), mundane verbs (*go*, 去). This means that the composition of automatically acquired Internet corpora reflects general language in a way similar to a manually constructed representative corpus.

The frequency profile across newswire corpora also reflects the difference between them: the *Reuters* corpus is known to have a large number of news items from the markets, while the *Xinhua* corpus contains mostly generic news. So financial terms are more significant in *Reuters*, while the *Xinhua* corpus has a higher frequency of place names and words used in the description of events, e.g. 记者 (journalist), 将 (is about to begin), 与 (participate).

Table 3 shows the most significant differences between the frequency lists of word forms in representative corpora vs. Internet corpora. In addition to the above mentioned technical reason (differences in lemmatisation) the use of lists of word forms helps as it reveals more facts concerning the use of specific forms, such as *Posted* (capitalised and in the past tense), which is an indicator of the time, when a message appeared on the Internet. The list of word forms also makes it clear that the BNC shows much greater use of past forms (*was, had, said*) and third person pronouns (*she, he, her,* it). This correlates with another study of the language use on the Web made by William Fletcher (2004), who also remarks that "the BNC data show a distinct tendency toward third person, past tense and narrative style, while the Web corpus prefers first and second person, present and future tense and interactive style".[5]

Words that are more frequent in the BNC include several interjections (*er, Yeah, Oh*), which typically occur in fiction stories, as their authors use them to imitate spoken language. As discussed earlier, fiction is underrepresented on the Internet, while the language of chat rooms makes very little use of hesitation markers such as *er.*

As for the *Russian Reference Corpus*, the comparison shows limitations of its pilot version. First, like newswire corpora, the RRC overuses a word referring to a particular news source, in this case the *Izvestia* newspaper (like *dpa* or *Xinhua*), which is explained by the fact that its newspaper component is to a significant degree composed of texts from *Izvestia*. Second, it contains several large texts such as the Russian Criminal Code, which is responsible for the high frequency of *Российской Федерации* (Russian Federation), *Кодекс* (legal code), *Статья* (an article in a legal code). The RRC also contains several fiction stories, which are responsible for the high frequency of words referring to their main characters (*Danilov, Jonas, Mitya*).

It is not surprising that words more frequent in Internet corpora include Internet-specific words (*Web, site, email*) or words related to interaction with it (*Click, program, Reply*), as well as words referring to hot topics at the time of corpus collection (*Bush, Yushchenko*). At the same time the differences between word frequencies in the Internet and representative corpora are much less significant than those for corpora based on newswires. Finally, even though the BNC and I-EN are roughly the same size (100 million words), the number of different words in the Internet corpus is much larger, leading to a token-type ratio of 143 in the BNC vs. 63 in the English Internet corpus (Table 4, see Appendix). This means that the Internet corpus covers more topics, hence giving a broader sample of language use.

4. Related research

The Internet provides a readily available resource for obtaining texts in a wide variety of languages. It is not surprising that it was used in many projects in corpus linguistics. See an overview in Kilgarriff and Grefenstette (2003). In this section we analyse a few approaches that directly concern the topic of text collection on the Internet and identify the advantages of the methodology proposed in the current study.

As mentioned in the introduction, Kilgarriff (2001) introduced the notion of the Linguistic Search Engine, which would use a list of selected URLs to create a virtual corpus, which should be distributed over servers. If a page from the list is not available at the time of querying, it can be replaced by any other page with similar characteristics (following the same methods as used by Google in their "Show similar pages" link). Unfortunately this approach has not been put into practice, probably because of the inherent difficulty involved in maintaining and querying a distributed corpus. Later on, the same idea was used by Oxford University Press (Kilgarriff, personal communication) for development of a new Internet-based representative corpus for English that should replace the BNC in dictionary development within OUP. However, the results of this project are not available for the academic world and are restricted to English only.

Resnik and his colleagues (Resnik & Smith 2003) proposed a new approach, STRAND, which assumes mining the Web for collecting parallel corpora. They used the Altavista search engine to locate candidate sites that most probably contain translated pages (for instance in English and Chinese) on the basis of their structural similarity. More recent versions of STRAND have a spider component, which traverses the web starting from pages with identified

parallel texts to following links to other pages, which might also contain parallel texts. The databases for parallel texts in several languages with download tools are available from the STRAND webpage. Recently they also applied the same technique for collecting a set of links to monolingual pages identified as Russian by http://www.archive.org, an Internet archiving service. We have evaluated the Russian database produced by this method and identified a number of serious problems with it. First, it does not identify the time when the page was downloaded and stored in the Internet archive. Given the dynamic nature of the Internet, many webpages and even servers become obsolete in a few years. Out of a sample of 10,000 URLs from the database, almost a quarter (2,483) of URLs link to nonexistent pages or servers. Second, the STRAND database of Russian links contains many pages in other languages that also use Cyrillic script (typically Bulgarian and Ukrainian). Google can also wrongly attribute the language of a page, but a combination of its detector with 4 keywords frequent in a given language effectively eliminates the possibility of retrieving a wrong page. Finally and most importantly, the STRAND database contains many links to pages that do not contain any stretches of connected text, such as databases of photos, catalogues of online shops or dating services. At the same time, all pages in our corpora contain connected text.

The same problem of retrieving pages with connected text appeared in a study by Fletcher (2004), who collected an Internet corpus by making queries to the Altavista search engine. He ran a series of queries for the ten highest frequency words in the BNC (such as *the, of*), retrieved a corpus about 7,000 documents (after filtering duplicates) and reviewed all of them manually. In the result he selected 5,000 documents with a reasonable amount of connected text (i.e. he discarded about 30% of documents) following the estimation of Ide et al. (2002) for the minimum of 2000 words as an indicator of connected text. Even though no manual review was possible for our collections of 40,000–50,000 documents, manual coding of metatextual properties of about 200 documents for each language discovered no documents without long stretches of connected text. The mean document length in the English Internet Corpus is 3006 words per document (after filtering navigation frames out). As mentioned above, longer queries and the use of common words provide sufficient safeguards against including documents with high level of noise.

Baroni and Bernardini (2004) developed BootCat, a tool for downloading webpages through the Google API, and used it for creating domain-specific corpora by means of bootstrapping. In the first step, a small set of seed words is identified for a domain and used for collecting the first version of a corpus from the Internet. In the second step, they compute the frequency profile of the

first corpus against a reference corpus using the log-odds ratio and produce a new set of keywords for collecting a larger corpus. Our methodology is based on some of their tools, but it is designed for collecting new reference corpora by using words from the general lexicon.

Probably the earliest study on how to measure the relative frequency of lexical items was the comparison of frequency lists from the *Brown* corpus against the LOB corpus (Hofland & Johansson 1982), which used the chi-squared test (introduced by Pearson for testing the independence of two variables). After that work there was a considerable discussion on the applicability of the chi-squared test in corpus studies. One of the problems is that the chi-squared test relies on the assumption of the normal distribution of word frequencies in texts, while words in natural language are not distributed at random. Once a rare word like *whelk* or *Noriega* occurs in a text once, it is likely that it will be repeated several times (Church 2000). So Dunning (1993) proposed the use of log-likelihood statistics (originally for detecting collocations), and Kilgarriff (1997) used the Mann-Whitney test. Rayson and Garside (2000) argue that log-likelihood statistics provides the most reliable method for highlighting words more specific for a given corpus. This is why it is also used in this study.

5. Conclusions and further perspectives

The set of successfully downloaded URLs from Step 3 constitutes an "open-source" corpus, i.e. a corpus that exists as a list of URLs with additional open-source software for downloading the set of HTML pages and post-processing them (such as removing navigation frames, tables, duplicate pages, etc). This means that a large corpus can be recreated from the list of URLs in simple steps that do not require human intervention.

The most formidable problem in corpus development is how to make the corpus available to other researchers. Kilgarriff (2001) refers to copyright as "the hobgoblin of corpus builders". You cannot distribute texts from your corpus unless you have explicit consent of the copyright holder to do so. In some cases the permission is given on the webpage itself, e.g. the GNU Free documentation licence, but in most cases you have to request it, which is quite time consuming for a list of 40,000 URLs. At the same time the current copyright law allows downloading a publicly accessible webpage for your own personal use (assuming that it is its function to be read by Internet users).

Providing URL links to Internet pages is also subject to legal regulation. There have been several claims aimed at prohibiting other companies making

"deep links" to any interior page of their websites. Isenberg (2000) analyses a case of this sort, which was *Ticketmaster vs. Tickets.com*. In the ruling the judge stated: "hyperlinking does not itself involve a violation of the Copyright Act (whatever it may do for other claims) since no copying is involved". However, in another clause the ruling restricts the use of "deep linking" with respect to the concept of fair competition, allowing it in cases when "deep linking" does not create a mistaken identity for the second site. The same reasoning of "fair competition" was used in a similar case (*Danish Newspaper Publishers Association* vs. *Newsbooster*) held in Copenhagen to make just the opposite ruling that prohibited Danish company Newsbooster from providing deep links to websites of Danish newspapers on the ground that services offered by Newsbooster are in direct competition with similar services of the content providers. Given that the list of links in an Internet corpus does not give the impression that websites it points to are the same as the URL list itself nor compete with services provided by respective websites, a corpus in the form of lists of URLs is not subject to copyright restrictions.

The Internet allows easy creation of large corpora: anyone can use this methodology to create a corpus they want. However, it is not enough to compile a new corpus for one's own study. There is a general need for reference corpora that establish a benchmark and can be reused in other research done on the same language, so that the results are comparable and reproducible. For instance, two studies of expressions of futurity can be compared by reference to the same material. In terms of reproducibility, everyone can check the frequency of *strong tea* in the BNC and study the list of collocates for *strong*. An automatically acquired corpus can be considered as a reference only if it is documented with respect to its content.

The proposed methodology offers two ways of doing this: by describing the corpus macrostructure, i.e. its composition, and its microstructure, i.e. specific properties of its lexicon. In terms of the *macrostructure*, we can take a sample of documents from the corpus and describe it using a small set of metatextual categories corresponding to mainly external classification criteria (Sinclair 2003). Experience shows that it takes less than 2 minutes on average to describe a text. This adds a one person-day contribution for coding a statistically acceptable sample of 200 documents from a new Internet corpus.

The *microstructure* of a newly acquired corpus can be described by comparing its frequency list against the frequency list of another corpus for the same language, if it is available. This will reveal information about lexical variation between the two corpora. The analysis of the corpus microstructure can be also extended to the comparison of bigrams, POS tags or other statistical properties.

The comparison of automatically acquired corpora to large representative and newswire corpora shows that Internet corpora are similar in their macro- and microstructure to their more representative counterparts: the English Internet corpus shares similar features with the BNC, when they are both compared against the *Reuters* corpus. Obviously, the composition of Internet corpora is much more varied than that of newswire corpora: they include many texts written from the personal viewpoint, expressing one's attitudes, covering a large variety of topics, instructing or entertaining the reader, making recommendations, etc. This is also reflected in the lexicon: Internet corpora use many more sentences with personal pronouns, question words, simple action verbs, etc. This means that in the absence of representative corpora, which are rare and expensive to build, Internet corpora can provide a better window into general use of modern language than widely available newswire corpora.

In some respects, Internet corpora can be even more useful than manually constructed representative corpora. Firstly, they present a sample of language as it is used now, not the language of 1980s, as in the BNC (cf. the example of *browser,* discussed above). Secondly, the design of manually constructed corpora raises awkward theoretical questions about the sampling techniques used to determine their representativeness. What is the reason for the large amount of fiction and social science texts in the BNC in comparison to natural science and technology? Fiction frequently uses language in an odd way, which does not represent adequately its every-day use, especially if we consider a text like "FNS" from the BNC (this is "Alice in Wonderland", which also does not fit into the implied timeframe). Texts available on the Internet have their own peculiarities: medical and legal advice and texts on computer science topics are particularly frequent, whereas fiction is rare for copyright reasons. Nevertheless, an Internet corpus is representative for the language the Internet users are exposed to.

The methodology opens up the possibility of creating freely-available large reference corpora for a variety of languages using very limited resources. Such reference corpora can be used by other researchers in the same way as the BNC is used as a benchmark for making their own lexicographic studies, assuming that the corpus contains a snapshot of the language used in webpages at the time of its collection. More and more pages in time will become unavailable. This will gradually reduce the value of the URL list, though the corpus collection procedure can be repeated at regular intervals (every two-three years) to update the list (and consequently the description of corpus composition). URL lists for corpora described in the paper and software tools used for collecting and processing them are available for download from http://corpus.leeds.ac.uk/internet.html.

Finally, let us return to the original example with the collocation *strong tea* and the possibility of detecting it in a corpus. Internet corpora collected in the reported experiment are large enough to detect such collocations as *strong tea* (32 uses in I-EN, log-like score 22.19), *крепкий чай* (112 uses in I-RU, log-like score 219.59) and 浓茶 (29 uses in I-ZH, log-like score 63.28), whereas the German corpus does not contain a significant collocation for *Tee* (tea) corresponding to *strong tea*. German dictionaries list two potential translations *starker Tee* and *kräftiger Tee*, but there are very few examples in the corpus (3 in total), which gives a hint that those collocations are not as salient in German as in English, Russian or Chinese. This can be confirmed by searching the Internet. Google reports finding about 2,000 pages on the German Internet for all forms of *starke*[*rmn*] *Tee* and *kräftige*[*rmn*] *Tee*, whereas for Russian it finds 32,000 pages for the two most frequent forms (*крепкий чай, крепкого чая*) alone, not counting the eight other forms. Google also gives 32,400 hits for *strong tea* and 71,000 for 浓茶.

This example shows that corpora collected in the experiment are much smaller than the total Internet content for respective languages, and they cannot answer all questions that are of interest for lexicographers in the same way as the BNC cannot. However, Internet corpora can answer many more questions than Internet search alone. First, they can be searched using linguistic criteria, at least with respect to lemmas and POS tags, e.g., to study the difference between *suggest*+VBG vs. *suggest*+*to*+VB. Second, it is possible to produce sets of collocates, e.g. the English Internet corpus for *strong* gives *partnership, communication skills, family*, none of which are considered as significant by the BNC. It is also possible to obtain other statistical data from Internet corpora, such as word and n-gram frequencies, whereas the Internet search is restricted to approximate "Number of pages", which also cannot be trusted. For instance, a study of the Google output by Veronis (2005) shows that a search like (Chirac OR Sarkozy) produces *less* results than a search for a single term in the OR expression.

In short, Internet corpora are easy to collect and easy to distribute, they are comparable to large representative corpora in their size and composition, they can serve as a benchmark for comparing results of independent studies, and they are more useful for linguistic purposes than Internet search engines.

Acknowledgments

I am grateful to comments on the earlier drafts of the paper made by Judy Delin, Tony Hartley, John Sinclair, Martin Thomas and the two anonymous reviewers. Special thanks to Marco Baroni, who released the BootCat tool and made this type of research widely available.

Notes

1. It is interesting that *powerful tea* also occurs in the BNC, but only 3 times in a single source (text EES), which is exactly about corpus linguistics and collocations, e.g.: *"strong" and "powerful" may be similar, but we can not talk of "a strong car" or "powerful tea"*. This raises the question about the possibility to use in corpora linguistics texts referring to examples that are often unnatural or claimed to be impossible.

2. The choice of the name here is misleading (Gigaword), as Arabic and Chinese corpora are smaller than 1 billion words (the Arabic corpus is about 350 MW, the Chinese corpus is of 1 billion characters, while most Chinese words consist of 2–3 characters). At the same time, the English Gigaword corpus is about 1.6 GW.

3. URL (Universal Resource Locator) is the path to an Internet page, typically starting with http:// or ftp://.

4. In parenthesis we list examples of subclasses of respective categories, which do not constitute a closed-class list, but can help in making the decision for classification of a page. Basic categories on the other hand do constitute a closed-class list to choose from.

5. Even though the first person pronoun *I* is in the list of words more frequent in the BNC, this reflects the fact that many Internet writers use the lower case *i* in this function.

References

Aston, G. & Burnard, L. (1998). *The BNC Handbook: Exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.

Baroni, M. & Bernardini, S. (2004). BootCaT: Bootstrapping corpora and terms from the web. In *Proceedings of the Fourth Language Resources and Evaluation Conference* (pp. 1313–1316). Lisbon, Portugal.

Berglund, Y. (2000). *Gonna* and *going to* in the spoken component of the British National Corpus. In C. Mair & M. Hundt (Eds.), *Corpus linguistics and linguistic theory* (pp. 35–49).

Church, K. (2000) Empirical estimates of adaptation: the chance of two Noriegas is closer to p/2 than p^2. In *Proceedings of the 17th conference on Computational linguistics* (pp. 180–186). Saarbrücken, Germany. (http://acl.ldc.upenn.edu/C/C00/C00–1027.pdf)

Čermák, F. & Křen, M. (2005). Large Corpora, Lexical Frequencies and Coverage of Texts. In *Proceedings from the Corpus Linguistics Conference Series*, Vol. 1, no. 1. Birmingham, UK, July, 2005. (Available at: www.corpus.bham.ac.uk/PCLC)

Dunning, T. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19 (1), 61–74.

EAGLES (1996). *Preliminary recommendations on text typology.* EAGLES Document EAG-TCWG-TTYP/P. (http://www.ilc.cnr.it/EAGLES96/texttyp/texttyp.html)

Finn, A., Kushmerick, N. & Smyth, B. (2002). Genre classification and domain transfer for information filtering. In *Proceedings of the European Colloquium on Information Retrieval Research* (Glasgow 2002). (http://www.smi.ucd.ie/hyppia/publications/ECIR02/)

Fletcher, W. (2004). Making the Web more useful as a source for linguistic corpora. In U. Connor & T. Upton (Eds.), *Corpus Linguistics in North America 2002: Selections from the Fourth North American Symposium of the American Association for Applied Corpus Linguistics* (pp. 191–205). Amsterdam: Rodopi. (Also http://miniappolis.com/KWiC-Finder/)

Halliday, M.A.K. (1966). Lexis as a linguistic level. In C. E. Bazell, J. C. Catford, M. A. K. Halliday & R. H. Robins (Eds.), *In Memory of J. R. Firth* (pp. 148–162). London: Longman.

Hofland, K. & Johansson, S. (1982). *Word frequencies in British and American English*. Harlow: Longman.

Ide, N., Reppen, R. & Suderman, K. (2002). The American National Corpus: More Than the Web Can Provide. In *Proceedings of the Third Language Resources and Evaluation Conference* (LREC) (pp. 839–44). Las Palmas, Spain.

Isenberg, D. (2000). *Ticketmaster v. Tickets.com*. (Available from: http://www.gigalaw.com/library/ticketmaster-tickets–2000–08–10-p1.html)

Kilgarriff, A. (1997). Using word frequency lists to measure corpus homogeneity and similarity between corpora. In J. Zhou & K. Church (Eds.), *Proceedings of the 5th ACL workshop on very large corpora* (pp. 231–245). Beijing and Hong Kong.

Kilgarriff, A., (2001). The web as corpus. In P. Rayson et al. (Eds.), *Proceedings of Corpus Linguistics Conference 2001*. Lancaster. (Available from http://www.itri.bton.ac.uk/techreports/ITRI–01–14.abs.html)

Kilgarriff, A. & Grefenstette, G. (2003) Introduction to the Special Issue on the Web as Corpus. *Computational Linguistics*, 29 (3), 333–347

Lee, D. (2001). Genres, registers, text types, domains, and styles: clarifying the concepts and navigating a path through the BNC jungle. *Language Learning & Technology,* 5 (3), 37–72. (http://llt.msu.edu/vol5num3/pdf/lee.pdf)

O'Donnell, M. (1995) From Corpus to Codings: Semi-Automating the Acquisition of Linguistic Features. In *Proceedings of the AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*, Stanford University, California, March 27–29, 1995. (Also http://www.wagsoft.com/Coder/)

Rayson, P. & Garside, R. (2000). Comparing corpora using frequency profiling. In A. Kilgariff & T. B. Sardinha (Eds.), *Proceedings of the Comparing Corpora Workshop*, 38th ACL 2000 (pp. 1–6). Hong Kong.

Resnik, P. & Smith, N. A. (2003). The Web as a parallel corpus. *Computational Linguistics*, 29 (3), 349–380. (Also http://www.umiacs.umd.edu/~resnik/strand/)

Sharoff, S. (2004). Methods and tools for development of the Russian Reference Corpus. In D. Archer, A. Wilson & P. Rayson (Eds.), *Corpus Linguistics Around the World* (pp. 167–180). Amsterdam: Rodopi.

Sinclair, J. M. (2003) Corpora for lexicography. In P. van Sterkenberg (Ed.), *A Practical Guide to Lexicography* (pp. 167–178). Amsterdam: Benjamins.

Szmrecsanyi, B. (2003). Be Going to Versus Will/Shall: Does Syntax Matter? *Journal of English Linguistics*, 31 (4), 295–323

Upton, G. & Cook, I. (2001). *Introducing Statistics*. Oxford: Oxford University Press.

Veronis, J. (2005). Web: Google's missing pages: mystery solved? (A discussion available from http://aixtal.blogspot.com/2005/02/web-googles-missing-pages-mystery.html)

Author's address

Serge Sharoff
Centre for Translation Studies,
School of Modern Languages and Cultures,
University of Leeds
e-mail: s.sharoff@leeds.ac.uk
tel: +44–113–343 7287
fax: +44–113–343 3287

Appendix

Table 1. The balance of text types in collected corpora

		BNC	I-EN	RRC	I-RU	I-DE
Authorship	Corporate	18%	44%	–	38%	51%
	Male	28%	23%	50%	18%	13%
	Female	13%	3%	25%	6%	2%
	Unknown	4%	11%	16%	15%	14%
	Multiple	36%	19%	9%	23%	20%
Mode	Written	90%	86%	100%	84%	90%
	Electronic	0%	13%	0%	16%	9%
	Spoken	10%	1%	0%	0%	1%
Audience	General	27%	33%	–	40%	61%
	Informed	47%	45%	–	46%	31%
	Professional	26%	22%	–	14%	8%
Aim	Discussion	–	45%	–	47%	45%
	Information	–	11%	–	4%	25%
	Recommendation	–	34%	–	35%	21%
	Instruction	–	6%	–	3%	5%
	Recreation	–	4%	–	11%	4%
Domain	Life	27%	14%	51%	25%	12%
	Politics	19%	12%	18%	10%	21%
	Business	8%	13%	3%	7%	5%
	Natsci	4%	3%	2%	3%	1%
	Appsci	7%	29%	3%	19%	18%
	Socsci	17%	16%	16%	5%	8%
	Arts	7%	2%	6%	2%	4%
	Leisure	11%	11%	1%	26%	31%

Table 2. Words less/more frequent in news corpora

Words more frequent in the BNC		vs. Words more frequent in Reuters	
Lemma	**LL-score**	**Lemma**	**LL-score**
you	6005.14	*say*	8559.54
I	5271.42	*percent*	4513.35
she	3334.57	*million*	2364.29
be	2411.89	*market*	1982.47
do	1610.71	*billion*	1518.25
they	1502.79	*bank*	1468.84
your	1282.15	*company*	1258.34
can	1191.74	*newsroom*	1240.37
what	1090.53	*share*	1214.84
my	1023.56	*tuesday*	1199.25

Words more frequent in I-EN		vs. Words more frequent in Reuters	
Lemma	**LL-score**	**Lemma**	**LL-score**
you	4343.16	*say*	12154.94
I	2797.67	*percent*	3424.40
your	2731.17	*million*	2103.23
or	1845.60	*market*	1943.17
my	1262.80	*bank*	1574.68
can	965.08	*billion*	1270.30
this	899.29	*newsroom*	1254.03
use	729.11	*share*	1193.56
me	719.46	*its*	1175.01
do	687.78	*company*	1125.64

Words more frequent in I-DE			vs. Words more frequent in IDS		
Word form	**Gloss**	**LLscore**	**Word form**	**Gloss**	**LLscore**
ich	I	1227.77	*Mark*	Mark	858.82
dass	that (new)	691.60	*Uhr*	hour	528.01
mir	me-dat	350.78	*Prozent*	percent	329.20
du	you-fam	376.29	*daß*	that (old)	307.32
mich	me-accus	273.24	*sei*	be-conjunct	291.95
the	–	266.27	*dpa*	dpa	262.05
Ich	I	250.70	*bis*	to	258.87
Du	You-fam	241.12	*Millionen*	millions	235.37
of	–	198.39	*gestern*	yesterday	225.47
Beiträge	contributions	178.55	*SPD*	SPD	181.97
Beitrag	contribution	155.29	*sagt*	said	177.19

Words more frequent in I-ZH			vs. Words more frequent in Xinhua		
Word	**Gloss**	**LL-score**	**Word**	**Gloss**	**LL-score**
我	I	12994.63	日	day	3161.93
你	you	6238.64	将	be about to	3060.79
她	she	2648.87	新华社	Xinhua	2118.91
了	le particle	2387.15	经济	economy	2059.03
去	to go	1688.94	记者	journalist	2052.17
看	to see	1620.84	表示	say/inform	1966.47
那	that	1610.86	与	participate	1546.89
什么	what	1492.52	和	and	1510.03
着	zhe particle	1398.92	地区	region	1415.81
不	not	1159.02	台湾	Taiwan	1310.50

Table 3. Words less/more frequent in Internet corpora

Word forms more frequent in the BNC		Word forms more frequent in I-EN	
Word form	**LL-score**	**Word form**	**LL-score**
was	1251.29	*your*	303.43
had	953.62	*Posted*	278.37
he	928.66	*Web*	262.23
she	912.82	*program*	255.15
er	909.30	*Internet*	228.45
her	795.37	*site*	217.36
Yeah	623.65	*Click*	201.91
it	580.80	*Center*	192.76
erm	578.10	*online*	189.36
his	496.03	*Bush*	177.53
I	415.54	*email*	177.42
said	398.64	*information*	174.04
Oh	385.29	*New*	168.38

Word forms more frequent in the RRC			Word forms more frequent in I-RU		
Word form	**Gloss**	**LLscore**	**Word form**	**Gloss**	**LLscore**
Федерации	Federation	104.56	*это*	this	165.81
Российской	Russian	95.80	*вы*	you	137.88
Статья	Article	69.44	*что*	that	132.85
Данилов	Danilov	67.64	*i*	(translit)	122.34
Известиям	Izvestia-dat	50.09	*Вы*	You	118.33
Иона	Jonas	38.17	*сайта*	site-gen	109.59
Митя	Mitya	35.84	*для*	for	107.26
Андреевна	Andreevna	31.70	*Ответить*	Reply	90.51
Кодекса	legal code	29.99	*подробнее*	more	88.55
Дмитриевич	Dmitrievich	28.84	*Если*	If	74.06
Ивановна	Ivanovna	25.32	*Ющенко*	Yuschenko	69.06
стихи	verses	24.88	*я*	I	67.93
Известий	Izvestia-gen	22.25	*Рейтинг*	Page rank	62.45

Table 4. The size of Internet corpora

	I-EN	**I-DE**	**I-RU**	**I-ZH**
Number of tokens	126,643,151	126,117,984	156,534,391	91,385,075
Number of types	2,003,056	3,384,491	2,036,503	457,557
Number of documents	42,133	31,195	33,811	30,148
Average document length	3006	4043	4630	3031