Creating general-purpose corpora using automated search engine queries

Serge Sharoff

Centre for Translation Studies, University of Leeds
s.sharoff@leeds.ac.uk

1 Introduction

The Internet is a natural source of linguistic data, providing an abundance of texts of various types in a large number of languages. These texts are already in electronic form suitable for corpus studies, either as downloadable pages, or as a resource to be searched using search engines. On the other hand, large representative corpora of the size of the British National Corpus, BNC (Aston and Burnard, 1998) exist for very few languages, because they are expensive to build. They are absent even for major world languages, such as Chinese or French. Many ad-hoc text collections are available, but they are restricted in either their size or the variety of text types. Typically they are produced on the basis of out of copyright fiction (such as Project Gutenberg) or newswire/newspaper texts that are available in large quantities and relatively easy to acquire from their publishers, e.g. the Reuters corpus for English (Rose and Stevenson, 2002), or the "Gigaword" corpora for Arabic, Chinese and English (Cieri and Liberman, 2002). News corpora are useful for many applications, such as development of gazeteers, parsing and word sense disambiguation algorithms, yet they cannot replace corpora representative for general language, such as the BNC, as the former reflect only the formal register of reporting news stories, while corpora that are claimed to be representative should include a variety of text types. Below we compare the language of news corpora against the language used in the BNC and the language derived from the Web. The comparison shows that the news corpora differ significantly from either representative or Internet corpora and cannot provide a window into modern language use in general.

The usefulness of Web data is evidenced by numerous recent corpus studies based on the number of pages returned by Google for specific queries, see many references to this research in Section 2 of Kilgarriff and Grefenstette (2003). Some researchers in traditional linguistics also use data returned from Google as the basis for their research, cf. Robb (2003), Volk (2002). However, Google is a poor concordancer. It provides only limited context for results of queries, cannot be used for linguistically complex queries, such as searching for lemmas (as opposed to word forms), restricting the part of speech or specifying the distance between components in the query in less than crude ways. More importantly results are ordered according to their "relevance" to the topic of the query using page-rank considerations, not according to left or right context as it is often useful for corpus work. When two linguistic phenomena are compared on the basis on the number of results returned by Google, the counts cannot be trusted. For instance, Veronis (2005) analyses problems with the logic of Google output and shows (among other things) that a search like (Chirac OR Sarkozy) produces *fewer* results than a search for a single term in the OR expression.

The problems with ordering the results and the amount of returned contexts have been addressed by several projects, such as KWICFinder (Fletcher, 2004) or WebCorp (Renouf, 2003), which rely on Altavista or Google queries, but present results in the form of traditional concordances. However, this does not solve the problems with counts, query language and richer linguistic information.

The ideal solution for corpus linguists would be a Google-like engine adapted to linguistic criteria. Kilgarriff (2001) discussed this idea under the name of D3CI (Distributed Data Distributed Collection Initiative),

which would crawl the Web, collect a list of URLs to create a virtual corpus, which should be distributed over original 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 difficulties 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. Similarly, Marco Baroni and his colleagues (personal communication) recently started crawling the Web to collect large corpora for English, German and Italian.

A simpler methodology that does not involve crawling can be based on collecting a list of URLs from the Internet using the existing crawl index of search engines. For instance, Phil Resnik and his colleagues (Resnik and Smith, 2003) extended their technique for developing parallel corpora to collect a list of URLs of Russian pages from the web archiving engine `http://www.archive.org` . However, their list contains links to many pages that no longer exist or to pages that do not contain instances of connected text, such as price lists, collections of photos, etc. The same problem of retrieving pages with connected text appeared in a study by Fletcher (2004), who collected an Internet corpus by making a series of queries for the ten highest frequency words in the BNC, retrieved a corpus of 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. A similar technique was also used in CorpusBuilder (Ghani et al., 2003), though they did not evaluate the composition of their results and even give very little information about the size of their corpora. Baroni and Bernardini (2004) developed BootCat, a tool for downloading webpages through the Google API and applied it to creating specialised corpora. Further, Ueyama and Baroni (2005) used the tool for creating a general-purpose Japanese Web corpus of approximately 3.5 million words using query words taken from an elementary Japanese language textbook.

However, these experiments were not aimed at using Internet for building a BNC-like corpus, i.e. a corpus of at least 100 million words covering a variety of text types and domains. The aim of this paper is twofold. First, I investigate the possibility to develop a BNC-like corpus for a number of different languages (Chinese, English, German, Romanian, Ukrainian and Russian). Second, I present an evaluation of the collected corpora using the composition of resulted corpora and their frequency lists for some of the languages (English, German and Russian). Since large balanced corpora are available for English and Russian, we can compare our Internet corpus against their content. For English we use the BNC, for Russian — the Russian Reference Corpus (RRC), its pilot version used in this study contains about 35 million words, 45% of which is fiction, the rest is split between newspapers and various domains; for more information cf. Sharoff (2004).

2 DIY manual for a BNC

The method for collection of a large corpus for language X is based on BootCat (Baroni and Bernardini, 2004) and comprises four basic steps:

1. word selection: choose 500 word forms that are frequent in language X;
2. query generation: produce 5,000-8,000 queries, each of which contains 4 words from the word list from Step 1
3. downloading: send the queries to a search engine and collect the top 10 URLs returned for each query
4. post-processing: solve problems with encoding, boilerplate, duplicates

Now we will explain the rationale for the parameters used in each step.

2.1 Step 1: Word selection

Words in the query list should be sufficiently general, i.e. they should not indicate a specific topic. If a word like *Zeppelin* was used in the query list, this would create a bias in our corpus towards texts from the history of aviation or hard rock. On the other hand, function words frequently occur in pages that do not contain complete sentences, such as catalogues, captions for photos, price lists . For instance, *from* can bring a page from a holiday catalogue with a photo and caption: *Two weeks in Toscana, prices from 300 £*. If the goal of corpus collection is to provide examples of language use in connected texts, such pages should be avoided.

Many common frequent words indicate a particular topic, such as *work* or *room*. However, they can be used in the word list, as they do not bias the corpus because of their polysemy. Even when they are not polysemous, common words can still be used in frequency lists, if they indicate a large number of situations (see examples with *work* and *room* in Section 2.2 below).

Some studies, e.g. Kilgarriff and Grefenstette (2003); Ghani et al. (2003), considered the need to select words that are unique to the language of corpus collection. For instance, according to such views it is not advisable to use *restaurant*, as this word exists in several different languages. However, the query stage in the proposed methodology uses the language filter of a search engine, which by itself rarely makes mistakes in classification of pages. What is more, the presence in each query of three other frequent words in the target language eliminates pages in "wrong" languages.

Since general search engines (such as Google or Yahoo) do not perform lemmatisation, we have to rely on lists of word forms only. This can in principle distort results in the case of languages with elaborate morphology, such as Arabic, Romanian or Russian, in which a word may have 10-20 forms. Thus, a query based on exact word forms in such languages operates with words that are much rarer in comparison to English. For instance, two lemmas *high* and *высокий* are good translation equivalents having roughly the same rank and frequency in English and Russian, as their position in the respective frequency lists is around 180 and the frequency is around 500 instances per million words (ipm). However, the frequencies of the exact forms *high* and *высокий* are quite different: 290 ipm for *high* with the rank of 264 vs. 34 ipm for *высокий* with the rank of 2140 (the shift of its rank also reflects the number of forms of more frequent words). This means that for languages with rich morphology in the end we will find fewer pages, because in those languages we use less frequent word forms. Fortunately, this did not cause problems for our study, because many webpages exist in the languages under study (Romanian, Russian, Ukrainian) anyway, so we can find a sufficient number of hits for each query. At the same time, in languages with richer morphology it is possible to use only forms that are more likely to appear in connected text, such as verbs, because the presence of a verb indicates that there is a clause, while the presence of four verbs corresponding to a four-word query indicates a more elaborated description.

For English and Russian we used 500 frequent common words from the frequency lists from respective representative corpora. For English we used the frequency list of word forms collected by Adam Kilgarriff from the BNC. For Russian we used the frequency list of the RRC. For Chinese, German and Romanian we also started with frequency lists from existing corpora, which exhibited some bias towards news items. For Chinese it was the "Gigaword" corpus, consisting of Xinhua newswires (thus excluding the Taiwanese section of the "Gigaword" corpus, because it uses another version of Chinese characters). For German, the frequency list was based on the list of word forms from the IDS corpus from Institut für Deutsche Sprache. Even though the IDS corpus contains a variety of text types (including some fiction and texts from science and humanities), it is biased towards news sources. This is reflected in its frequency list: the word SPD (the name of a German political party) is more frequent in it than ja (yes), Kinder (children) or Frau (woman). We extracted from it the list of the most frequent 500 words which start with lower-case letters (adjectives, adverbs and verbs) and are not specific with respect to a topic, e.g. *häufiger* (more frequent), *wünscht* (wants), etc.

If we want to develop a corpus for a language and we do not have access to a frequency list, 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 (this was the case with the Internet corpus for Ukrainian).

We can use more words from the frequency list than the original suggestion of 500, however, this increases

efforts put into development of the query list (we spend more time cleaning the list from words we do not want) and increases the number of topic-specific words as we progress along the frequency list to less and less frequent words.

2.2 Step 2: Query generation

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 *n*oisy pages in the form of price lists, tables, lists of links, etc. Shorter queries and the use of function words result in more noise. Function words are invariably used in broken sentences, such as catalogues or lists of headlines, which are not ideal candidates for a corpus. The presence of one-two common words also does not guarantee an instance of connected text. For instance, the first page retuned by Google for the query *work* AND *room* includes several links to pages which do not contain stretches of connected text, such as `http://www.readingroom.com/aboutus/featuredwork.cfm`.

At the same time, a four-word query is much more likely to yield a page with narrative prose. For instance, the top ten pages produced by the query *work room hand possible* all have stretches of narrative prose ranging from two to five thousand words (not counting navigation frames). The pages retrieved also refer to a variety of domains, including a selection of summaries from Yahoo news, pages on political debates, orthopaedic surgery, forensic investigation analysis, classes offered in an art centre, a blog on maps, descriptions of furniture, electronic tools, fiction books and historical events. Even more specific words, such as *Scottish* in the context of a four-word query bring a variety of topics. For instance, the query *deep houses resources Scottish* returns pages devoted to history, architecture, politics, technology (production of energy), funding guides, etc.

However, if we use queries longer than four words, the number of pages returned gets 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 few hits 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. Even though there is sufficient amount of pages in Romanian, our task was to collect a corpus with proper encodings of diacritics, which are frequently omitted in Romanian Internet pages.

BootCat has a mechanism for automatic generation of a random list of N-tuples out of the original word list. In this experiment it has been extended with the mechanism of prefixing random stings with a specific string to achieve the following functionality. Search engines can restrict the search to a variety of languages using their own linguistic filters. However, 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 detecting Ukrainian we used the query *має* OR *її* (*has* OR *her*).

2.3 Step 3: Downloading

In the reported experiment we used the Google API (application program interface) via BootCat. Since then another API for Yahoo has been made available. For each query we take 10 top URLs returned by the Google API and use them for further processing. In the current setup we used 5,000 queries, which resulted in 50,000 URLs. However, some URLs can be found 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 to reach the target corpus size, say a corpus of more than 100 million words requires about 35-40,000 pages, given that downloaded pages contain on average about 3-4,000 words. The list of successfully downloaded URLs is stored in the corpus database and can be used to recreate the corpus by other researchers.

The procedure can be repeated to enlarge the corpus up to the limit of all texts in this language indexed by the search engine. However, a corpus of 100 million words gives abundant lexicographic data for words common in general language. According to our experiments with the languages under study, the top 25,000 words have at least 100 occurrences (words at the end of the 25000 word list in English include *exploitative,*

	I-EN	I-DE	I-RU
Number of tokens	126,643,151	126,117,984	156,534,391
Number of word forms	2,003,056	3,384,491	2,036,503
Number of lemmas	1,608,425	3,081,197	791,311
Number of URLs	42,133	31,195	33,811
Average document length (in words)	3006	4043	4630

Table 1: Some statistics for Internet corpora

lithograph, neutrophil, and some proper names). A concordance of 100 lines provides sufficient evidence for lexicographers, especially given that such words are typically monosemous, cf. the experience in development of the COBUILD dictionary (Sinclair, 1987). Words that do not provide this evidence in a 100 MW corpus (such as ones with 10 occurrences or less) are rare or misspelled words e.g. *oystercatcher* or *somtimes*. A study of terminology in the field of oystercatchers (a bird of the family of Haematopodidae) will require a specialised corpus.

The upper limit for an Internet corpus depends on what is a reasonable size for its storage and reasonable time for producing concordances. Currently the CorpusWorkbench, the tool we use for indexing and making queries, limits the size of annotated corpora (with POS and lemma tags) to about 200 million words. Some studies, e.g. Kilgarriff and Grefenstette (2003), show that many unsupervised algorithms (such as word sense disambiguation) steadily improve their performance on larger corpora reaching the size of one billion words. So for some applications it might be advisable to collect a larger corpus.

2.4 Step 4: Post-processing

Pages collected in the previous step are subjected to 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 use the lynx browser to convert pages from HTML into plain text. This works better than frequently used ad-hoc Perl filters, as it removes HTML add-ons, including javascripts or comments, but does not lose information on character encodings (lynx has options `display_charset` and `assume_local_charset` to render them correctly once we identified them for every page). Another advantage of Lynx is that after removing HTML tags it leaves traces of links in the original document, so that we can use simple heuristics to remove navigation frames (such as the density of links, which tend to appear mostly in navigation frames). Finally we can 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). The simplest procedure used for the Internet corpora reported in the paper involved detection of exact duplicates only. Since then, Baroni and Zanchetta [1] produced a tool for detection of shared n-grams in large text collections, which helps in finding near duplicates using the shingling algorithm (Broder et al., 1997): if several identical n-grams appear in two documents, this is an indication that the two documents share significant part of their text.

This sequence of steps results in a clean corpus in plain text format using a single chosen encoding. Finally, in order to create a proper corpus out of this collection of plain texts, we need language-dependent morphosyntactic processing, such as tokenisation (more important for Chinese and other languages without explicit word boundaries), lemmatisation (especially for morphologically rich languages), as well as part-of-speech tagging.

A summary of the characteristics of the Internet corpora collected for English, German and Russian is given in Table 1 (abbreviated as I-EN, I-DE and I-RU respectively). The size of corpora varies slightly: the longest pages have been retrieved for Russian, so the Russian corpus is slightly bigger. The most significant difference is in the number of lemmas in the lexicon: 791,311 in I-RU vs. 3,384,491 in I-DE. This depends partly on the features of a particular language and partly on properties of tokenisers and lemmatisers. For

[1] `http://wacky.sslmit.unibo.it/`

instance, in German there are many compound nouns, which in other languages are typically decomposed into several words, e.g. *Fachhochschulratspräsident* (the president of the council of polytechnic universities). This increases the amount of separate forms and lemmas. The smaller number of lemmas in Russian can be partly explained by the larger number of word forms per lemma, as well by more aggressive splitting done by the Russian lemmatiser used in the experiment (mystem[2]), which treats hyphens as word separators. In contrast, our English and German lemmatisers (respective versions of the TreeTagger[3]) treat the hyphen as a word character.

3 What is under the hood?

In this section we use two methods to compare Internet corpora against standard manually-collected corpora such as the BNC, Reuters or Gigaword. The first method involves assessment of corpus composition using a text typology, which is similar enough to the one used in the BNC to allow comparison between the BNC and Internet corpora. The second methodology involves comparison of lists of the most frequent words taken from various corpora to show the most significant differences in their lexicon.

3.1 Composition assessment

The reported procedure produces a corpus of about 40,000 texts, which is not practical to assess in its entirety, so we have to choose a representative sample. The issue of the representativeness of a text collection in terms of the number of documents is frequently neglected in corpus studies, whereas statistics offers a straightforward procedure to estimate the symmetric confidence interval, which is frequently used for determining the size of a sample required in sociological studies or polls:

$$\sigma = \pm c\sqrt{\frac{p(1-p)}{N}} \text{ (Upton and Cook, 2001, 301)}$$

where c is the percentage point (or the critical value) from the standard normal distribution appropriate to attain the desired confidence level, p is the estimated probability of an event, and N is the population sample required for the result to be within the given confidence interval with the given confidence. Note that the value of the interval does not depend on the size of the population. The only assumption is that the total population is significantly larger than the size of the sample. The confidence level refers to the probability that the real distribution measured on the complete population will be indeed within the symmetric interval. For the same sample size N we can make a statement with confidence of 90% ($c = 1.645$) or 95% ($c = 1.96$), giving a slightly larger symmetric confidence interval in the second case.

The total *population* in the case of a sociological study refers to the total number of people or cases which constitute the subject of the study, such as the number of voters in a country, while the *sample* refers to the focus group the study is based on. In the case of corpus studies, an Internet corpus is itself a sample of the population, i.e. the content of the Internet for a particular language, which in its turn is a sample of the total language used in the society. However, in terms of statistical analysis of its composition, the Internet corpus of 40,000 documents represents the total population, from which we take a sample in the form of a subset of URLs.

Application of the above formula is based on two assumptions: the normality of the sample distribution and the approximation of the probability of an option. The first assumption is justified by random sampling from a much larger list. The second assumption involves replacement of the unknown value of p, the probability of an option, e.g. the proportion of texts written by men, with its estimation from the number of options in the respective category. Categories in the text typology described below have 3-8 options, so we can estimate p as $0.125 \leq p \leq 0.33$. Of course, we cannot always make the assumption that all options in a category have equal probability. However, the value of $p(1-p)$ does not vary much: for any $0 \leq p \leq 1$

[2] `http://corpora.narod.ru/mystem/mystem.html`
[3] `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`

it is always true that $p(1-p) \leq 0.25$ and it gets smaller for smaller values of p providing a more precise symmetric interval.

In short, this means that if we take a random sample of 200 documents from a text collection, we can achieve the confidence interval of $\sigma = \pm 5\%$ and confidence level 90%. A better approximation of the corpus composition within the interval of $\pm 1\%$ with 95% confidence will require a much larger sample, of about 1,500 documents. In our experiments we used samples consisting of 200 documents, so the figures reported below assume the confidence interval of $\sigma = \pm 5\%$ with confidence level 90%.

3.1.1 Text typology and detection criteria

Assessment of the corpus composition requires a text typology to annotate texts in the sample. Existing research in corpus studies has produced two theoretically sound text typologies. First, an extensive text typology has been developed for coding texts 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. Second, the European Advisory Group on Language Engineering Standards (EAGLES) produced text typology guidelines in work headed by John Sinclair (EAGLES, 1996; Sinclair, 2003). The EAGLES guidelines include functional categories, however, they do not cover many text types that are frequent in general-purpose corpora or webpages, such as types of newspaper texts or advertisements. Finally, 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 and the results will be less reliable.

We attempted to develop a small set of categories and rules for assigning values to 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, so that each document requires no more than 5–8 choices from the list of categories. The coding itself was done using the Systemic Coder (O'Donnell, 1995), which provides an interface for prompting choices for each text and allows basic statistical analysis of the results.

Another requirement for the set of categories is the reliability of information provided in Internet pages for detecting their values. For instance, the gender 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 this cannot be done, either because it is ambiguous, like *Chris* in English, or the sex association is not known to the coder, as is the case with *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, a guess about the age of the author or the size of the intended audience is much less reliable, so these were not included in the classification scheme.

We assess each text using 5 categories: authorship, mode (aka channel), knowledge expected from the audience, the aim of text production and the generalised domain. The basic set and the order of categories follows the EAGLES guidelines and corresponds to the degree of certainty in coding values of those categories: it is quite easy to code the authorship, while many texts cover several domains at once, so the choice of the domain is less reliable. In order to reduce possible ambiguity in choosing the values of categories we provide explicit instructions for filling their values on the basis of observable features of texts. In a trial study four colleagues were asked to code a sample of 100 texts according to the proposed typology. They all completed the task in less than an hour with very small variation in the set of assigned categories.

Full results of assessment of the composition of automatically acquired corpora are shown in Table 2.[4] The English and Russian Internet corpora can also be compared against data obtained from representative corpora for those languages, though the comparison cannot be complete, as neither the BNC nor the RRC classify pages with respect to the purpose of their production. The *audience level* code from the BNC cannot be directly compared against the knowledge expected from the audience according to our typology, while in the RRC there is no coding for this category at all.

[4]Since additional annotators did not assess the complete sample, the results listed in the table are based on my own counts.

In the following subsections we describe the set of categories in detail and give instructions for making decisions about choosing their values.

3.1.2 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 (such as references to author's husband, third person pronouns referring to the author, grammatical agreement, etc).
- **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; this applies to texts created by government or non-profit organisations as well). There can be some inconsistency here: a newsitem in the newspaper can lack the name of its author, while a feature article, which still carries a corporate copyright statement, can have an explicit author's name. In the latter case, the decision should be made for the single named author. On the other hand, a letter for investors has been claimed to be written by the CEO of a company, but since it represents the position of the company and most probably it was edited by the whole board of directors (if not external consultants), it should be coded as corporate. The same applies to such documents as Papal Encyclicas or declarations in the name of the heads of governments.
- **unknown** – no information about the author is available on the page nor can it be inferred without significant extra efforts.

The result of codings reported in Table 2 show that Internet corpora in comparison to traditional representative 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).

3.1.3 Mode

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

- **written** – traditional written texts, including online newspapers, homepages, etc;
- **spoken** – transcripts of sound-wave recordings, including interviews;
- **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.

		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: Comparison of corpus composition

3.1.4 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 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** – no knowledge about the topic is required for reading this text, e.g. a text on ulcers from the BBC website. Such texts are written for the broadest general public. They refrain from using terminology that the general public is not expected to know.
- **informed** – some general knowledge of the topic is required, e.g. a description of ulcers for medical students. Another example could be an explanation of the design of home theatres for audiophiles Such texts are not very technical, but they do use a significant amount of specialist terminology.
- **professional** – significant prior knowledge about the domain is required for reading a text, e.g. an article in the Journal of Gastroenterology and Hepatology. Such texts are written for professionals using many abbreviations, dense terminology, etc. They also appear on specialised websites. This does not assume that the category is limited only to topics from respected professions. A discussion of the number of *ingots for GM tinkering* in "Ultima Online" is classified as aimed at the professional audience as well.

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. The instruction for coders states

> If you can easily understand the text content, choose **general**; if you can in principle understand what the text is about, but it contains special terminology, choose **informed**; if you cannot understand the text, choose **professional** (if you are a specialist in the domain of the text, try to imagine yourself to be a layman)

In terms of their composition, Internet corpora contain a good balance of 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.

3.1.5 Aims of text production

This is the classification of texts according to their function in the society, as borrowed from Sinclair (2003), but with some modifications outlined below:

- **discussion** – texts aimed at discussing a state of affairs (e.g. articles in newspapers, academic papers, travel stories).
- **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. This category is not necessarily concerned with leisure activities (cf. the subtypes of text domains discussed below).

- **instruction** – such texts are aimed at educating their readers; the following subclasses 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 example in 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; this is the most discoursive type of instructive texts).

- **information** – texts whose primary purpose consists in providing information. Sinclair (2003) restricts the category to reference compendia, but in corpora we find many other cases, such 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**). Note that this category is limited to texts only concerned with data dissemination. A discussion of the history of the Tory party in the Wikipedia is classified as **information**, while the Tory election manifesto is **recommendation**.

There are some borderline 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 to become a potential customer or supporter, it is classified as **recommendation**, a text without obvious propaganda is **discussion**. If the tests for other categories do not produce convincing results, the general rule for coding text production is to choose **discussion**.

A classification of this sort is used neither in the BNC nor in the 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, services or political movements).

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 the Internet corpora). Texts aimed at recreation are more frequent in I-RU (11%), including OCR'd versions of fiction texts and exchanges of jokes, but still they are relatively rare.

3.1.6 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 studies of terminology and corpora and refer to the unsuitability of "trying to arrange a hierarchy of simple topic labels". However, in practical terms the offered list of some 30 domains is too fine-grained. What is more, a webpage can be a subject to far 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.

- **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. **life** incorporates fiction (imaginative texts in the BNC), weblogs on dating or parenting of a child; **world affairs** from the BNC is treated as **politics**. 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.

There are fewer texts from arts, humanities and social sciences 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 such subdomains as computer science, medicine or construction industry).

If we compare this data against the Retuers corpus (a newswire corpus annotated with domain codes), we will find that 56% of the Reuters corpus consists of financial news (its C, E and M subcategories), contrasting with 13% of business texts in the Internet corpus (8% in the BNC). At the same time less than 0.5% of texts in the Reuters corpus is classified as science (GSCI), which includes the **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 aimed at giving information in the form of news reports. This suggests again that Internet corpora can be claimed as more representative as newswire corpora, such as Reuters or Gigaword.

3.2 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. The comparison of frequency lists is a much faster way of understanding the major differences between the newly acquired corpus and a known benchmark corpus and judging how significant they are. Also unlike the corpus composition exercise, which starts with a predefined set of categories, comparison of frequency list is driven exclusively by data found in corpora (even though it is influenced by the results of tokenisation and lemmatisation).

Among various methods for comparing frequency lists we choose the log-likelihood statistic, following the reasoning that it provides the most reliable method for comparing frequency lists (Rayson and Garside, 2000).

The computation of the log-likelihood statistic is based on the following contingency table:

	Corpus 1	Corpus 2	Total
Frequency of word	a	b	a+b
Frequency of other words	c-a	d-b	c+d-a-b
Corpus size	c	d	c+d

Then the expected values E1 and E2 and the log-likelihood value G2 are calculated as:

$$G2 = 2(a\ln(\tfrac{a}{E1}) + b\ln(\tfrac{b}{E2})); E1 = c\tfrac{a+b}{c+d}; E2 = d\tfrac{a+b}{c+d}$$

In the study reported below we calculated log-likelihood values for the frequency of lemmas or word forms in two 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. The analysis should highlight statistically significant differences between the frequency lists and can suggest ways in which one corpus is less balanced than the other. 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.

More in BNC	LL-score	More in Reuters	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

Table 3: BNC vs. Reuters

More in I-EN	LL-score	More in Reuters	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

Table 4: I-EN vs. Reuters

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 (Table 3).

Second, we compare an Internet corpus against a newswire corpus with known composition (the English Internet corpus against the Reuters). 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 (because of relatively high frequency of hits from newspapers in concordance lines). 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 (Tables 4 and 5).

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-likelihood scores are shown in Table 6. 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 automatically acquired Internet corpora. This boosts differences in lemma lists significantly without any underlying linguistic reason.

Tables 3 and 4 show that newswire corpora in comparison to both the Internet and the BNC overuse words referring to financial data (*million, Mark*), specific entities and institutions (*market, dpa*), 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*). Another specific feature of newswires is much greater use of reported speech, which is reflected in the overuse of such words as *say, sagen.* 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

More frequent in I-DE			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-subjunc	291.95
the	-	266.27	dpa	dap	262.05
Ich	I	250.70	bis	to-temporal	258.87
Du	You_{fam}	241.12	Millionen	millions	235.37
of	-	198.39	gestern	yesterday	225.47
Beiträge	messages	178.55	SPD	SPD	181.97
Beitrag	message	155.29	sagt	said	177.19

Table 5: Comparing I-DE vs. IDS corpus

More in BNC		More in I-EN	
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

Table 6: Comparing the BNC to I-EN

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.

At the same time words that are *less* frequently used in newswire corpora follow the same pattern as established by the difference between the Reuters corpus and the BNC. Newswire corpora in comparison to the BNC and Internet corpora use fewer first and second person pronouns, question words (*what, welche*), modals (*can, muss*), mundane verbs (*go, gehen*). This means that the composition of automatically acquired Internet corpora reflects general language in a way similar to a manually constructed representative corpus.

Finally, Table 6 shows the most significant differences between the frequency lists of word forms in representative corpora vs. Internet corpora. In addition to the abovementioned 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 used on the Web made by 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 frequently occur in transcripts in the spoken component of the BNC, as well as in fiction stories, as their authors use them to imitate spoken language. As discussed earlier, fiction is underrepresented in the Internet, while the language of chat rooms makes very little use of hesitation markers such as *er*.

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.

4 Conclusions and further research

The proposed procedure described in Section 2 is applicable to any language with more or less significant Internet presence. The procedure can produce a large corpus (100-200 million words) which, as shown in Section 3, can be considered as comparable to large representative corpora in terms of its size and coverage of various domains. What is more, the corpus can be considered as "open-source", as it exists as a set of URLs accompanied by additional open-source software for downloading the set of HTML pages and post-processing them (i.e., removing navigation frames, tables, duplicate pages, etc). If the parameters of an Internet corpus are described with adequate precision, it can function as a benchmark used by other researchers in the same way as in the BNC. For instance, everyone can use the BNC to compare the frequency of occurrences of *strong tea* and *powerful tea* and make conclusions about their most typical contexts, for instance, by referring to the fact that *powerful tea* occurs three times in a single text and the reason why it occurs there is that that text is exactly on the topic of corpus linguistics and collocations and the example is used to illustrate collocations impossible in English.

If we claim that an Internet corpus is useful as a benchmark for studying language X, it is necessary to understand how stable the benchmark is. If you do your study for English on the basis of the BNC, there can be minor variations depending on the version of the BNC you are using. However, changes concern a tiny portion of the whole corpus: the number of occurrences of *powerful tea* will not change. Kilgarriff

[5]There may be several reasons why the first person pronoun *I* is in the list of words more frequent in the BNC. One possybility is that many Internet writers use the lower case *i* in this function.

(2001) defended the possibility to use Internet corpora, which are dependent on the transient nature of the Internet, by referring to a scientific study of water taken from river Lune: you cannot expect that molecules are exactly the same, yet the study is replicable. However, chemical analysis provides ways for measuring how replicable the study is.

If we distribute an Internet corpus in the form of URL lists, one possible measure can concern the half-life of those lists, i.e. we can measure how many URLs from the original list are still accessible after a certain period and how much of the content of the respective pages is the same. Research in this area is still in its infancy, so we would like to study it more closely in collaboration with Marco Baroni. The parameters for studying the URL half-life will include the number of URLs retained, the proportion of the text remained exactly identical, differences in the frequency of retrieved words, differences between URL sets for languages.

In addition to studying changes in Internet corpora derived from a fixed set of URLs, one can study variations caused by differences in the collection procedure. Ueyama and Baroni (2005) conducted a study of two Japanese corpora collected according to the same procedure using the same list of query words, but the first corpus was collected in July 2004, the second one in April 2005. The study shows that the composition of the two corpora varies considerably (even the intersection between the two sets of URLs is below 20%). It would be interesting to extend this research by studying the rate of change of web-derived corpora using the influence of several other parameters apart from the variation in time, such as the differences:

- between languages and cultures: all languages exhibit explosive grow on the Internet, but one can expect that the rate of change for languages currently less present on the Internet is more significant;
- between search engines: we can study the difference between corpora derived using Google, Yahoo or our own crawling engines (also using different methods of crawling);
- between sets of query words selected from various sources (such as frequency lists) according to the same procedure (such the one outlined in Step 1 below)
- between procedures for selecting query words: we can also study the difference between corpora produced using function words, adjectives only, words specific to a domain (e.g. set of headwords from Encyclopedia Britannica), etc.

More in I-EN2	LL-score	More in I-EN1	LL-score
I	143.14	tea	70.47
June	120.60	Christmas	34.21
Posted	99.64	dog	27.17
book	62.09	and	24.01
Definitions	51.45	Tea	22.37
blog	50.74	Speaker	21.00
that	47.98	PST	20.34
think	47.02	Feb	20.21
References	45.66	dogs	19.46

Table 7: Comparing two Internet corpora collected from different query words

As for now we can briefly show preliminary results for the URL half-life and corpus variation. We created two English Internet corpora collected respectively in February and June, 2005, using two sets of 500 query words without any intersection between the two sets. Both lists were extracted from the BNC frequency list. The June list included the most frequent words, e.g. *chance, minutes, simple, thank*, while the February list consisted of less frequent common words, e.g. *opinion, purpose, suddenly, unemployed*.

An experiment in August, 2005 involved downloading a random selection of 1000 links from each of them. 934 links from the February corpus and 982 links from the July corpus were still available. Further experiments are necessary for determining the rate of degradation. As for the difference caused by sets of

query words, Table 7 shows the comparison between the frequency lists of the February corpus (I-EN1) and the June one (I-EN2). The differences (measured by the log-likelihood score) are much less significant in comparison to what has been reported in Tables 4 and 6.

An interface to Chinese, English, German and Russian corpora, respective URL lists, lists of queries and the results of corpus assessment are available from `http://corpus.leeds.ac.uk/internet.html`.

Acknowledgements

I'm grateful to Marco Baroni for useful discussions and for releasing BootCat, the tool that made this type of research widely available.

References

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

Baroni, M. and S. Bernardini. 2004. Bootcat: Bootstrapping corpora and terms from the web. In *Proc. of the Fourth Language Resources and Evaluation Conference*, Lisbon, Portugal.

Broder, Andrei Z., Steven C. Glassman, Mark S. Manasse, and Geoffrey Zweig. 1997. Syntactic clustering of the Web. In *Proc. Sixth International World-Wide Web Conference.*

Cieri, C. and M. Liberman. 2002. Language resources creation and distribution at the linguistic data consortium. In *Proc. of Language Resources and Evaluation Conference (LREC02)*, pages 1327–1333, May. Las Palmas, Spain.

EAGLES. 1996. Preliminary recommendations on text typology. Technical Report EAGLES Document EAG-TCWG-TTYP/P, EAGLES.

Fletcher, W. 2004. Making the web more useful as a source for linguistic corpora. In U. Connor and T. Upton, editors, *Corpus Linguistics in North America 2002: Selections from the Fourth North American Symposium of the American Association for Applied Corpus Linguistics.*

Ghani, R., R. Jones, and D. Mladenic. 2003. Building minority language corpora by learning to generate web search queries. *Knowledge and Information Systems*, 7(1):56–83.

Ide, N., R. Reppen, and K. Suderman. 2002. The american national corpus: More than the web can provide. In *Proc. of the Third Language Resources and Evaluation Conference*, pages 839–844, Las Palmas, Spain.

Kilgarriff, A. 2001. The web as corpus. In *Proc. of Corpus Linguistics 2001*, Lancaster.

Kilgarriff, A. and G. Grefenstette. 2003. Introduction to the special issue on the web as corpus. *Computational Linguistics*, 29(3):333–347.

O'Donnell, M. 1995. From corpus to codings: Semi-automating the acquisition of linguistic features. In *Proc. of the AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*, Stanford University, California, March 1995.

Rayson, P. and R. Garside. 2000. Comparing corpora using frequency profiling. In *Proc. of the Comparing Corpora Workshop at ACL 2000*, pages 1–6, Hong Kong.

Renouf, Antoinette. 2003. Webcorp: providing a renewable data source for corpus linguists. *Language and Computers*, 48(1):39–58.

Resnik, P. and N.A. Smith. 2003. The web as a parallel corpus. *Computational Linguistics*, 29(3):349–380.

Robb, Thomas. 2003. Google as a quick 'n' dirty corpus tool. *Teaching English as a Second or Foreign Language*, 7(2).

Rose, T. and M. Stevenson, M. amd Whitehead. 2002. The reuters corpus volume 1: from yesterday's news to tomorrow's language resources. In *Proc. of the Third International Conference on Language Resources and Evaluation*, Las Palmas de Gran Canaria.

Sharoff, Serge. 2004. Methods and tools for development of the Russian Reference Corpus. In D. Archer, A. Wilson, and P. Rayson, editors, *Corpus Linguistics Around the World*. Rodopi, Amsterdam.

Sinclair, J.M. 2003. Corpora for lexicography. In P. van Sterkenberg, editor, *A Practical Guide to Lexicography*. Benjamins, Amsterdam, pages 167–178.

Sinclair, John M., editor. 1987. *Looking up: an account of the COBUILD Project in lexical computing*. Collins, London and Glasgow.

Ueyama, M. and M. Baroni. 2005. Automated construction and evaluation of a japanese web-based reference corpus. In *Proceedings of Corpus Linguistics 2005*, Birmingham, UK, July.

Upton, Graham and Ian Cook. 2001. *Introducing Statitstics*. Oxford Univeristy Press, Oxford, 2 edition.

Veronis, J. 2005. Web: Google's missing pages: mystery solved? May.

Volk, M. 2002. Using the web as a corpus for linguistic research. In R. Pajusalu and T. Hennoste, editors, *Tähendusepüüdja. Catcher of the Meaning. A festschrift for Professor Haldur Õim*. University of Tartu.