A Uniform Interface to Large-Scale Linguistic Resources

Serge Sharoff

Centre for Translation Studies
School of Modern Languages and Cultures
University of Leeds, Leeds, LS2 9JT, UK
s.sharoff@leeds.ac.uk

Abstract

In the paper we address two practical problems concerning the use of corpora in translation studies. The first stems from the limited resources available for targeted languages and genres within languages, whereas translation researchers and students need: sufficiently large modern corpora, either reflecting general language or specific to a problem domain. The second problem concerns the lack of a uniform interface for accessing the resources, even when they exist. We deal with the first problem by developing a framework for semi-automatic acquisition of large corpora from the Internet for the languages relevant for our research and training needs. We outline the methodology used and discuss the composition of Internet-derived corpora. We deal with the second problem by developing a uniform interface to our corpora. In addition to standard options for choosing corpora and sorting concordance lines, the interface can compute the list of collocations and filter the results according to user-specified patterns in order to detect language-specific syntactic structures.

1. Introduction

Two types of practical problems confront research in translation studies and the training of translation students. The first stems from the limited resources available for targeted languages and genres within languages. Translation researchers and students are frequently interested in novel uses of words, e.g. *unhinged* in the sense of 'crazy', terminology in a specific domain, e.g. *peer-to-peer network*, or uses of moderately frequent words, e.g. *integrity* or *combat*, which exhibit significant polysemy and whose translations depend on the context of their use, none of which are adequately represented in bilingual dictionaries.

Such interests demand: sufficiently large corpora representative of modern language (at least of the size of the BNC) or large corpora that are specific to a problem domain. The requirement for large corpora stems from the Zipfian distribution of word frequencies. For instance, *integrity* occurs just 10 times in one million words of the Brown Corpus vs. 1467 in 100 million words of the BNC. The situation with collocations is much worse, as the Brown corpus has no instances of *to undermine one's integrity*, and a single instance of *to question one's integrity* in the form of *if his integrity were questioned*, which does not help in deciding whether it represents a recurrent pattern or whether the two words can be combined in any other ways. A BNC-like corpus is much better in this respect, but building such corpora is an expensive and time-consuming enterprise, so they are not available for many languages.

The requirement for modern corpora stems from the need to reflect recent trends in language use, because translators most typically work with modern texts. For instance, they might be interested in finding verbs most frequently co-occurring with *roadmap* to compare them against Russian verbs co-occurring with дорожная карта, the corresponding expression in Russian.

However, very few corpora can actually satisfy these demands, even for major European languages. The BNC is the best available resource, but it is slightly outdated, as it reflects the language of 1970s, 80s and early 90s. It is unlikely that the expensive procedure for making a representative corpus of modern British English will be repeated again in the near future. French lacks an equivalent of the BNC, and even French newspaper corpora (primarily, Le Mond) are not publicly available. For German there are interfaces to two corpora: a huge (1 billion words) corpus of IDS (Institut für Deutsche Sprache, though heavily biased towards newspapers) and DWDS, a BNC equivalent of 100 million words. However, both interfaces are limited with respect to options for doing searches and the corpora are not available as files in the same way as the BNC is available for English. The situation is much worse for lesser-studied languages, such as Polish or Thai, for which corpora are very scarce. However, in a multilingual research and teaching environment it is necessary to deal with such languages on demand.

Finally, translators nearly always specialise in specific domains, so they need corpus data coming from sources within their domain. In this respect the BNC is too general, as it contains relatively few technical texts. The total size of texts from all domains of engineering in the BNC is about 1.7 million words (in terms of size this is only marginally better than the Brown Corpus). Also these texts are almost exclusively from the domains of computing and electronics. And even they are severely outdated, and are not helpful for making modern translations. For instance, it does not contain the word *browser* in the sense of 'web-browser', because it was compiled before the Internet era.

The second problem concerns the lack of a uniform interface for accessing the resources, even when they exist. The possibilities for creating concordances, studying collocations or gathering statistical information vary quite significantly. For instance, it is difficult to compare patterns of uses of *anger* in the BNC and *Ärger* in the German IDS corpus, when the German interface uses the log-likelihood score for listing collocations, while the two options for the BNC are MI-score and z-score. What is worse, the German corpus in comparison to the BNC is heavily biased towards newspapers, so the data is not comparable anyway, as references to emotions in the IDS corpus are comparatively rare.

Corpora also vary in the query languages they use: an interface to each corpus uses its own query language. Compare, for instance, the query languages used in SARA (XAIRA), the IDS corpus interface and CorpusWorkbench (which is used as the interface to a variety of corpora), especially if you want to search for something that is beyond a simple word form search, e.g. uses of *bring about* with several intervening words, such as the *individuals who brought changes about* or *it was your efforts that brought that conference about.* Given that the background of translators is typically in the humanities, very few of them are proficient in query formalisms, and the need to learn several query languages and several interfaces in order to work with their corpora makes those resources much less usable.

This requirement is especially important in the classroom context for training future translators. It is not practical to teach corpus methods to translators using different interfaces and query languages. Also an exercise for a group of students should be studied using comparable corpora.

2. Acquisition of Corpora from the Internet

2.1. The procedure

We deal with the corpus-availability problem by developing a framework for semi-automatic acquisition of large corpora from the Internet for the languages we need. The framework is based on BootCat (Baroni and Bernardini, 2004), a suite of tools for automating queries to search engines. The methodology for corpus acquisition involves:

1. developing a list of words that are frequent in a specific domain (for a domain-specific corpus) or in language in general (for a BNC-like corpus);
2. creating a list of queries that randomly combine several words from the query list;
3. sending the query list to a search engine and downloading pages from the set of URLs returned as the result of queries;
4. post-processing the downloaded pages;
5. analysing corpus composition

To retrieve a general language corpus we can use 300-500 words. The best practice is to use word forms from the top of a frequency list for a language, removing function words and words denoting a specific topic. For instance, *conditions, clearly, ground* all are good candidates for the common word list, as they do not refer to a specific domain. On the other hand, for a domain-specific corpus it is reasonable to use a shorter list of query words for key concepts that identify the domain, e.g. *consolidants, deterioration,* or *gilded* are good words for collecting a corpus in the domain of artwork restoration and preservation.

The length of the list of queries produced in the second step corresponds to the size of corpus we would like to have. To reach the target of 100 million words, a general language corpus needs 5,000-8,000 queries. A domain-specific corpus combines fewer keywords and can be built using 100-1000 queries. The use of four common words in a query brings pages that contain relatively long pieces of connected text, unlike price lists, tables, lists of links, etc. The use of three-four domain-specific words also helps in better identification of the domain. A search for *deterioration* brings pages with references to degenerative dementia, the security situation in Iraq and problems with data reading from CD-ROMs. However, a search for *consolidants* AND *deterioration* AND *gilded* brings a variety of pages exactly on the desired topic.

The goal of the fourth step is to remove navigation frames and duplicates and to perform basic tokenisation, tagging and lemmatisation, if tools are available for a specific language. Tokenisation is necessary for Oriental languages (e.g. Chinese and Thai), while lemmatisation is especially important for highly inflected languages (e.g. Polish and Russian). The methodology for corpus collection is described in greater detail in (Sharoff, 2006).

2.2. Results

To cater for the needs of our researchers and students we developed general Internet corpora for a range of languages, including Chinese, French, German, Italian, Polish, Romanian, Russian and Spanish, as well as several more specific corpora, covering the domains of computer science and artwork restoration. The size of the general corpora ranges between 100 and 200 million words, which make them suitable for lexicographic research. Internet corpora also better reflect recent trends in modern language use. For instance, the BNC contains 37 instances of *roadmap*, none of which is used in the sense of a political plan (almost all examples in the BNC are from the computer domain), while the English Internet corpus contains 276 examples in a variety of uses, such as: *The factors we used to measure financial fitness provide a solid* ***roadmap*** *for local policymakers to take steps to improve and enhance each region's economic strength.*

We also attempted to assess the composition of English, German and Russian general-language corpora (I-EN, I-DE and I-RU), by coding their samples of 200 documents using a principled set of text description categories (Sharoff, 2004), which combine the experience of coding the BNC (Lee, 2001) and suggestions from (Sinclair, 2003). The set of categories used for composition assessment includes authorship (such as male, female or corporate), mode (written, spoken transcript), the audience (general, informed or professional), the aim of text production (e.g. discussion, recommendation or instruction) and the generalised domain (e.g. sciences, humanities or politics).

The results of this study show that, contrary to the popular belief that the Internet consists mostly of pornography and advertising cf. also (Crystal, 2001; Volk, 2002), the corpus of web pages created by this method contains a wide variety of topics and text types. The results for these three languages show that even though Internet corpora are not completely identical (some of them contain more texts of specific types than others), they are nevertheless comparable in the same was as the BNC is comparable to the German DWDS corpus, so that lexical patterns can be studied and compared across languages.

Internet-corpora are similar to the BNC in the proportion of texts produced by institutions and privately, even though

the number of private texts from female authors on the Internet is relatively small. The figures are consistent for the three languages studied: 3-6% of explicitly named female authors vs. 20-30% for men. Internet corpora also contain texts written for a variety of purposes: texts aimed at discussing a state of affairs, encyclopaedic entries and reports, instructive texts (manuals and tutorials), etc. The only significant difference concerns fiction, which is treated as an important category in traditional corpora, but fiction texts are relatively rare on the Internet; in I-EN and I-DE they constitute just 3-4% (11% in I-RU).

The most significant difference between I-EN and the BNC concerns the amount of texts from arts and humanities vs. those from sciences. 24% of the BNC consists of texts classified as socsci and arts. The amount of such texts in I-EN 17% looks similar, but 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 sciences in the Internet corpus: 7% in the BNC vs. 29% in I-EN (also in a variety of domains).

We also collected a set of domain-specific comparable corpora using seed words that are more or less exact translation equivalents. A query combining three-four terms from a problem domain identifies a topic, for instance:

En: *autosave, configuring, debugger, user-friendly*

Es: *autoguardar, configurar, depurador, amigable*

Ru: автосохранение, настройка, отладчик, дружественный

Zh: *zìdòngbǎocún, pèizhì, diàoshì, yǒuhǎo jièmiàn*

If we generate 500 queries of this type, we will be able to collect relatively large comparable corpora for these languages consisting of about 15 million words each. Corpora produced by these queries also show a good balance of text types, e.g. descriptions of features of products, manuals and FAQs, overviews in magazine articles.

3. A Uniform Interface

We dealt with the second problem by developing a uniform interface to our corpora. Standalone concordancers, such as MonoConc or Wordsmith, are not efficient for processing very large corpora (of the size of 100-200 million words) with morphosyntactic markup (at least with POS tagging and lemmatisation). They also do not cope well with the variety of encodings used in non-European languages (Arabic, Chinese, Russian, Thai). Finally, their options for making complex queries, such as those involving free word order or discontinuous constituents, are limited. Corpus query engines, such as CWB, DWDS, IDS, SARA (XAIRA), Sketch Engine, are capable of dealing with large lemmatised corpora. Their query languages are also sufficient for making advanced queries, however, as discussed above, the range and complexity of query languages, as well as the diversity of interfaces hinders the use of these resources, especially in the classroom context.

These considerations led to development of a single online interface to all corpora (Figure 1), which contains standard options for choosing corpora, sorting concordance lines by the left and right context, as well as producing lists of collocations. To simplify understanding of the syntactic structure in foreign languages, there are options for highlighting POS tags in the concordance and for mapping words to bilingual dictionaries (available for some language combinations).

The corpus search engine used in the Leeds CQP interface is powered internally by the IMS CorpusWorkbench (Christ, 1994), which is capable of rapid retrieval of information from large corpora and has a powerful query language. It allows queries with regular expressions, free word order constituents, expressions combining POS and lemma restrictions, conditions on XML tags, etc. The corpus query language used by CWB suits advanced users. However, the interface offers an option of simple queries akin to Google. A simple query term corresponds to a lemma, while a term in double quotes corresponds to a word form.

To express the possibility of a distance between elements, the query can include the + sign followed by the maximal number of words that can occur in the gap:

$bring \ \ +2 \ \ about$

This translates into a CWB query:

[lemma='$bring$'] []{0,2} [lemma='$about$']

The notorious MU query syntax of CWB designed for making free word order queries can be simplified using the + sign followed by a lemma:

$+plant + environment + damage$

translates into:

MU(meet (meet [lemma='$plant$'] [lemma='$environment$'] s) [lemma='$damage$'] s)

Finally, the availability of language-specific POS tags in our corpora provides options for making complex queries addressing POS placeholders to compare, for instance, the number of cases when *suggest* is followed by a verb in the *-ing* form vs. *suggest* followed by infinitives. As the CWB query in this case is relatively complex and the exact names of tags should be memorised:

[lemma='$suggest$'] [pos='VBG']

[lemma='$suggest$'] [lemma='to'] [pos='VB']

we implemented a query builder for CWB expressions, which assists in choosing POS tags specific to a given language and adds them to the query to build a well-formed CWB expression.

More advanced options in the interface include detection of collocations using any combination of three well-defined scores: log-likelihood, MI and T scores (Manning and Schütze, 1999) and filtering of the results according to a specific pattern. The latter option is similar to the output of Word Sketch Engine (Kilgarriff et al., 2004), which structures the list of collocations according to the set of predefined relations. As our interface is designed to work for a variety of languages, the pattern filtering mechanism can be specified by the user and applied on the fly in order to detect language-specific syntactic structures. Even for English the Sketch Engine grammar does not cover all syntactic aspects that might be of interest for a translator, such as pronouns or verbs with clause complements, like in the two examples mentioned above *suggest+VBG* vs. *plan+to+VB*. It is much more likely that a predefined set of sketches for another language will miss information required by the translator.

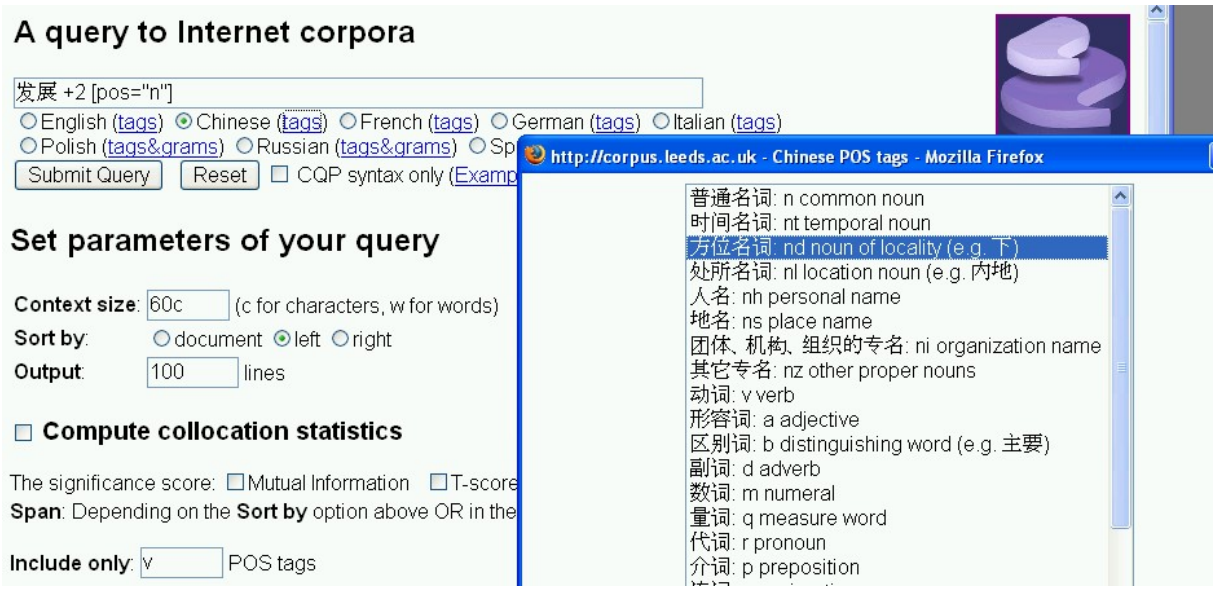

Figure 1: The query interface

Figure 2 shows the list of nouns which are common direct objects of the Chinese verb *fāzhǎn* (to develop). The list includes *qūshì* (trend, 234 instances), *qiánjǐng* (perspectives, 148 instances), *kōngjiān* (space, 211 instances), which is indicative of cases when *develop* is not a suitable translation equivalent for *fāzhǎn*. Other translation students working, for instance, with Italian and Russian can do the same exercise in the same interface to study conditions for translating *sviluppare* and развивать (the standard 'translation equivalents' of *develop* for Italian and Russian).

Collocation	Joint frq	Freqc 1	Freqc 2	LL score	
发展 ~~ 趋势	234	21861	1660	638.79	Show examples
发展 ~~ 经济	303	21861	16908	506.80	Show examples
发展 ~~ 战略	175	21861	2437	415.47	Show examples
发展 ~~ 前景	148	21861	952	411.70	Show examples
发展 ~~ 空间	208	21861	6780	403.70	Show examples
发展 ~~ 方向	211	21861	8335	389.14	Show examples
发展 ~~ 过程	206	21861	11352	345.62	Show examples
发展 ~~ 观	121	21861	3134	248.78	Show examples
发展 ~~ 阶段	133	21861	5057	247.72	Show examples
发展 ~~ 潜力	88	21861	842	226.07	Show examples
发展 ~~ 事业	124	21861	5198	224.87	Show examples

Figure 2: Filtered collocations

4. Conclusions

The corpora and query tools described above proved their usefulness in various activities in translation research and training. The interface is quite stable and will be released soon in open source together with the coming open-source version of the Corpus Workbench.

5. References

Marco Baroni and Silvia Bernardini. 2004. Bootcat: Bootstrapping corpora and terms from the web. In *Proc. of the Fourth Language Resources and Evaluation Conference, LREC2004*, Lisbon.

Oliver Christ. 1994. A modular and flexible architecture for an integrated corpus query system. In *COMPLEX'94*, Budapest.

David Crystal. 2001. *Language and the Internet*. Cambridge University Press, Cambridge.

Adam Kilgarriff, Pavel Rychly, Pavel Smrz, and David Tugwell. 2004. The Sketch Engine. In *Proceedings of Euralex 2004*, pages 105–116, Lorient, France.

David Lee. 2001. Genres, registers, text types, domains, and styles: clarifying the concepts and navigating a path through the BNC jungle. *Language Learning and Technology*, 5(3):37–72.

Christopher Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.

Serge Sharoff. 2004. Towards basic categories for describing properties of texts in a corpus. In *Proceedings of the Forth Language Resources and Evaluation Conference, LREC 2004*, Lisbon.

Serge Sharoff. 2006. Creating general-purpose corpora using automated search engine queries. In Marco Baroni and Silvia Bernardini, editors, *WaCky! Working papers on the Web as Corpus*. Gedit, Bologna.

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

Martin Volk. 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.