Concordances of Snippets

Elżbieta DURA
Lexware Labs
Göteborg
Sweden
elzbieta@lexwarelabs.com

Abstract

Excellent concordances can be produced by tools mounted on regular web search engines but these tools are not suitable for quick lookups on the web because it takes time to collect ad-hoc corpora with occurrences of a queried word or phrase. It is possible to get a web concordance in an instant if the amount of transferred data can be limited. One way to do it is to use snippets from a search engine as a basis for concordance lines, which is a solution adopted in Lexware Culler - a web concordance tool mounted on Google. It takes the same time to look up words and phrases in Lexware Culler as it takes for Google to deliver results for a search. The question is whether concordances based on snippets can be satisfactory for linguists or language learners. Our tests show that they actually can. With proper filtering concordances based on snippets can provide a good survey of current language use, which is particularly important as a complement to online dictionaries.

1 Introduction

The access to the web is ubiquitous, it is a self-renewing language resource and its size and variety exceeds all previous corpora. The counterpart of the public web was estimated up to 28 million books already in 2002, which can be compared with the largest number of volumes held by Harvard University - about 15 million (O'Neill, Lavoie, Bennett, 2003). Excellent concordances are produced by tools mounted on regular web search engines but these tools are not suitable for quick lookups on the web because it takes time to collect ad-hoc corpora with occurrences of a queried word or phrase. Is it possible to get a web concordance in an instant?

2 Web search engines in linguistic service

As the implementation of a special purpose linguistic search engine lingers on,[1] web search engines are used to produce web concordances. Search engines improve constantly. For instance it is no longer true that "Some search engines, including Google, FAST and Lycos, do not support wildcards at all" (Kehoe and Renouf, 2002). In Google wildcards are available for words and in AltaVista wildcards were available for both words and characters until its unfortunate recent death (1st April 2004). Google covers 4.28 billion web-pages, it has snap-shots of majority of them in its cache and its result lists include snippets - short text excerpts from matching web-pages showing a search term in its closest context. Google is not immaculate. A search term cannot be longer than 10 words and it is not always included in a snippet. It does not support case sensitive search or wildcards for characters.

2.1 Tools for concordancing the web

Concordance tools mounted on web search engines enable a user to compile own corpora from web-pages for a chosen search term and produce a concordance of the gathered text material.

2.1.1 Concordances collected in batch mode

KWiCFinder seems to have been the first one to provide linguists with KWIC concordances from the web. KWiCFinder is intended to be used in batch mode. It assists the user to formulate a query, the query is submitted to AltaVista, documents are retrieved and a KWIC concordance of 5-15 online documents per minute is produced. KWiCFinder is used in its own client application which needs to be downloaded.

2.1.2 Concordances from selected web-sites

WebConc is mounted on Google. It takes a search term from the user, accesses each web-page obtained from Google, collects all contexts of the search term and presents them as a concordance. It is perspicuous and easy to use. The maximal number of web-pages is limited to 50 in order to keep the retrieval time down but even with the minimum of 10 web-pages it is too slow for interactive use. It is possible to limit retrieval to

[1] There is actually one being currently developed and tested for English (http://lse.umiacs.umd.edu:8080/).

some chosen URL in WebConc, which probably is the best way to use the tool.

2.1.3 Concordances by e-mail

WebCorp (Kehoe and Renouf, 2002) makes access to each of web-pages retrieved by a chosen search engine, fortunately one does not have to wait for the results to appear on the screen because it is possible to order a concordance to be sent by e-mail. Various types of reports are made available, e.g. collocates of the search term can be presented summarized in a table. A frequency or alphabetically ordered list of all the words on any source page is available upon clicking on a URL link. Regular expressions can be used to express form alterations. WebCorp is an excellent example of how useful search engines can be made for linguists when their power is enhanced with natural language processing.

3 Instant web concordances

The web is not a true corpus: it is not representative of anything and it is not balanced. Nonetheless there is no better place to look up examples of current language use than the web, which possibly is also the most suitable type of use of this language resource. But interactive use, expected of concordances in general, requires short retrieval times.

3.1 Why web concordances are slow

WebCorp is said to be slow because “the current version of WebCorp is for demonstration purposes and the speed at which results are returned will increase as the tool is developed further”. Is the speed really in the hands of the developers of the system? The decisive factor here is the time it takes to access each web-page, which depends on the capacity of the data transmission channel and the actual server a web-page is on, and this is not even predictable. It is possible to make sure that a connection is always made to a quick server by using Google cache instead of original URLs but the time taken by data transmission still remains a problem. One possible solution is to rely on snippets for concordance lines. This saves the time needed to collect and transfer ad-hoc web corpora.

3.2 Web concordances and online dictionaries

The possibility to access current language use on the web in an instance is crucial as a complement to online dictionaries. The problem confronting dictionaries is how to handle two incompatible tasks simultaneously. One is to supply correct definitions and thereby preserve the usefulness of words. The other is to report on current trends in language usage, even when it means effacing meaningful differences between words. The role of an online dictionary complemented with concordances from the web would be to consider whether some popular usage may be based on confusion.

3.3 Lexware Culler

Lexware Culler builds concordances of Google snippets and it takes the same time to look up words and phrases in Lexware Culler as it takes for Google to deliver results. Language processing is applied not only to search terms but also to snippets from Google.[2] Besides Google wildcards which can be used for any word in general (*), it is possible to select words of a particular part of speech, in which case part of speech variables are used. Function word variables trigger expansion of search terms into alternative queries while variables of open parts of speech are used to filter away non-matching snippets obtained for a search term. This postfiltering is available for English and Swedish and it is being developed for Polish.

A table with a summary of results is always supplied in Culler along with concordance lines, which proves often handy, e.g. in investigation of collocations. It is possible to test the tool at http://82.182.103.45/lexware/concord/culler.html

3.4 Examples of use

Examples provided below are representative of the uses of Lexware Culler tested so far.

3.4.1 Context-based look-ups

It is not obvious how to find a word or a phrase in a dictionary if all one can go after is its context, it may be difficult even in a corpus unless very large. In order to find the word for a stick used in conducting an orchestra we made some futile checks in online dictionaries[3]. A simple query for “conductor’s *” in Culler yields *baton* directly with several examples like …*not the first soloist to feel the lure of the conductor’s baton*… .

A new adverbial use of the word *fett* (fat) has become very popular in the language of young Swedes – it has the role of a general magnifier. This use cannot be found in Swedish dictionaries or in the corpus of the Bank of Swedish. A typical context entered in Culler as a search term: “det är fett *” (it is fat *) gives 188 hits of which very few are examples of the basic uses of the word,

[2] Full-fledged language processing is available for Swedish. It is based on a language engine for Swedish – Lexware (Dura and Drejak, 2002).

[3] We tried WordNet, AskOxford.com, Merriam-Webster’s Collegiate Dictionary, Dictionary.com, and finally we found an example with *baton* in Cambridge Dictionaries Online.

majority of excerpts exemplify the new adverbial use.

A search term can be formulated as a typical defining context, for instance: “Moomin is a *” and “Moomins are *”. If the search is not limited to a specific country excerpts thus obtained are hard to find elsewhere side by side: *Many people in Japan think that Moomin is a hippopotamus, however, it is actually a forest fairy* or *Moomin is a Finnish cartoon/storybook character likened to Finland's version of Mickey Mouse*, *Moomins are WHITE, dammit!*.

3.4.2 Tracing mistakes, language changes

Do the French have the word *entrepreneur*? Yes, 233 000 French web-pages have “entrepreneur” or “entrepreneurs”. Has the correct spelling of Polish adverb *z powrotem* (back) lost to a new one word spelling *spowrotem* yet? Not, yet: it is used on 22 400 web-pages, while the correct one is used on 83 500 web-pages. Besides such simple checks Culler can be used to ferret out popular misinterpretations, such as those of the Swedish idiom *med berått mod* (in cold blood). Table 1 is a result summary for the search term: “med berått *” and “-mod” (any phrase beginning with *med berått* excluding those with *mod*).[4] The incorrect alternative versions thus extracted are: *mord* (murder), *lugn* (calm), *våld* (violence). The number of web-pages returned by Google is shown in the right column, the left one contains the number of concordance lines.

4	*med*	*berått*	*mot*	9
3	*med*	*berått*	*mord*	
1	*med*	*berått*	*lugn*	
1	*med*	*berått*	*våld*	

Table 1: Summary table for the search term “med berått *” “–mod”

3.4.3 Extracting unrestrained language use

One can learn from a dictionary what creatures typically produce grunting sounds. A further check in a large corpus yields yet more examples. The range of grunting creatures which appears in snippets for the search term “grunting like a *” and “-pig” is truly amazing: from more or less predictable ones, like *a walrus*, *a deranged gorilla*, *a wrestler*, *a lumberjack*, to rather unexpected ones, like *a constipated weasel*, *a freakin caveman*, *a eunuch impersonating Billy Idol*, plus many fresh associations like *an orc*, *the beasts back on Mordor*, etc.

In order to check whether and how expressions for becoming of age are dependent on the age itself the following queries were entered: “going on NUM” “gonna be NUM” “become NUM” “turn NUM”, “push NUM”, “reach NUM” “make it to NUM” “hit NUM”, where NUM stands for numerals. Enormous material was obtained for all ages, some of which involved surprises. For instance, “make it to NUM” had most hits in lower ages, where the lowest numbers referred mostly to the age of relations, while middle numbers referred to young people sick in some incurable illness; hitting 50 and more proved to be rare, probably because of its low news value.

4 Snippets as concordance lines

Whether snippets are sufficient as concordance lines is a question which can be settled empirically only. Culler has been used extensively for the past three months in uses for which the examples provided above are representative.

4.1 Google selections

An average snippet is about 20 words long, which in most cases is sufficient as disambiguating context. For each query Google retrieves max. 100 URLs and there may be up to 300 queries generated by Culler for a search term (when expanded with inflectional forms and/or function words).

Google selects web-pages according to a complex ranking, the main ingredient of which is the popularity of a web-page, measured among others in the number of links from other web-pages. Snippets are thus representative of prevalent language use. So are the numbers of matches reported by search engines because they report the number of web-pages with at least one match. The fact that each snippet is from a different web-page contributes to the diversity of excerpts.

The quality of excerpts can differ tremendously dependent on a search term. Generally the longer the search term the higher the chance for better excerpts. Some terms get snippets with proper name readings only, in which case it is better to limit the source of snippets to some large newspaper web-site[5].

4.2 Culler selections

Post-filtering of snippets is triggered either by variables of open parts of speech in a search term or by noise in excerpts. Three types of noisy excerpts are filtered away: repetitive, non-textual and non-phrasal. Each of the filters can be turned off. An average percentage of noise in snippets is

[4] Word filter in Google is applied to the whole web-page.

[5] The URL is entered in Culler’s slot for word filters.

about 20%.The three types of noises amount to an average of 21.7 % snippets discarded in the examples cited above (T is the number of snippets obtained from Google, D is the number of discarded snippets).

Search term	**T**	**D**	**%**
"Moomin is a *" "Moomins are *"	120	18	15.0
"grunting like DET *" "-pig"	236	28	11.9
"conductor's"	93	35	37.6
"z powrotem"	87	13	14.9
"spowrotem"	99	17	17.1
"entrepreneur"	183	49	26.7
"med berått *" "-mod"	9	0	0.0
"det är fett * * "	188	95	50.5

Table 2: Discarded snippets

4.2.1 World wide repetitions

Google does not return web-pages which are exact copies but it does return snippets which are the same or almost the same. Famous lyrics, dramas, stories, sermons, speeches, important news appear in enormous number of copies. For instance a search term "children at your feet" has about 50% repetitions, all of which involve web-pages with lyrics of the song "Lady Madonna". The average level of repetitive snippets is about 5% in the cited examples. Repetitive snippets discarded by Lexware Culler are the ones which differ only in:

- case of characters,
- date, time, link, and similar meta-information,
- word internal separators, like hyphens,
- language specific characters.

4.2.2 Non-textual snippets

Several types of more or less formulaic elements which are common on the web appear in snippets. None of these are usually desirable as concordance lines: boilerplate information, mathematical formulae, navigation tips, hyperlinks, e-mail addresses, post addresses, data on updates, headers, footers, copy right statements, logs, fragments of lists of items. 7% of snippets are discarded on average by this type of filtering in the cited examples.

4.2.3 Non-phrasal

Punctuation is ignored by Google while Culler departs from an assumption that phrasal context is normally requested, hence only snippets without interrupting punctuation within a search term are selected. Adding marginal wildcards to a search term is interpreted in Culler as a request for an unbroken phrasal context including words matched by wildcards. Snippets with search terms interrupted by commas, full-stops, colons, semicolons, question and exclamation marks are discarded by this filtering. The impact of this filtering differs very much from case to case: from half of the excerpts to none at all.

5 Conclusion

Instead of collecting ad-hoc corpora from each web-page retrieved by Google Lexware Culler builds concordances of snippets. Thanks to this limitation it is possible to look up words and phrases on the web in an instant. The quality of concordances built of snippets varies from excellent to poor dependent mainly on a search term but the goal of getting a quick glimpse of language use on the web is clearly attainable with snippets. Snippets are sufficiently long to provide disambiguating contexts. The ranking system of web search engines gives preference to the most popular web-pages, hence the prevalent language use can be expected in majority of excerpts. At the same time it is also true that extensive filtering is required in order to make accceptable concordance lines of snippets.

References

E. Dura and Marek Drejak. 2002. *Information Retrieval With Language Knowledge.* In: Cross Language Evaluation Forum Berlin Heidelberg New York: Springer-Verlag.

A. Kehoe & A. Renouf. 2002. *WebCorp: Applying the Web to Linguistics and Linguistics to the Web*. In: WWW2002 Conference, Honolulu, Hawaii.

KWiCFinder http://www.kwicfinder.com/KWiCFinder.html

E. T. O'Neill, B. F. Lavoie, R. Bennett. 2003. *Trends in the Evolution of the Public Web 1998 – 2002*. In: D-Lib Magazine, Vol. 9 Nr 4.

WebConc http://www.niederlandistik.fu-berlin.de/cgi-bin/web-conc.cgi?sprache=en&art=google

WebCorp http://www.webcorp.org.uk/