

Corpora, Serendipity & Advanced Search Techniques

Michael Wilkinson, University of Joensuu, Savonlinna Campus, Finland

ABSTRACT

Exploring corpora with concordancers can help translators to improve the quality of their translations by, for example, providing them with information about collocates; by helping them to choose between terms; or by enabling them to confirm intuitive decisions. But corpora also allow unpredictable, incidental learning: the user may notice unfamiliar uses in a concordance and follow them up by exploratory browsing. The article discusses the potential of corpora to throw up previously unknown information that may be relevant to a translation assignment, and illustrates how advanced search strategies can increase the likelihood of “accidentally” finding relevant information.

KEYWORDS

translation quality; corpora; corpus analysis tool; concordancer; key-word-in-context; serendipity; incidental learning; advanced searching

Introduction

As we know, there are known knowns; there are things we know we know.

We also know there are known unknowns; that is to say we know there are some things we do not know.

But there are also unknown unknowns – the ones we don't know we don't know.

– Thus spake Donald Rumsfeld, United States Secretary of Defence, in 2002 referring to the situation in Iraq, though he might well have been talking about searching corpora for translation candidates.

One way the translator can find out more about the “known unknowns” and the “unknown unknowns” is by exploratory browsing through relevant material. In this respect, a number of researchers in the fields of language learning and translating have drawn attention to the potential that electronic corpora, when used in conjunction with corpus analysis tools, provide for such “serendipitous” learning: corpora allow unpredictable, incidental learning in that the user may notice and explore unknown or unfamiliar uses in a concordance and go off at a tangent to follow them up.

In spring 2004, I began compiling a corpus of English-language tourism brochures with the aim of using it to teach students how the competent

use of electronic text corpora in conjunction with corpus analysis tools can help both the trainee translator and the professional translator to become better language service providers by enhancing both the quality of their work and their productivity, particularly when translating special field texts into a foreign language. (Many translators of non-literary texts in Finland frequently translate into their L2).

By September 2004, with the help of a student assistant, I had compiled a corpus amounting to 670,000 words. The Tourism Corpus contains mainly texts from brochures from the British Isles and from North America, especially Canada. When compiling the corpus, a major reason for including Canadian brochures was that they contain descriptions of activities that are often featured in Finnish source texts - e.g. snowshoe treks, skiing, snowmobile trips, wilderness adventures - which are rarely mentioned in British brochures. The file names were labelled with one of the following codes: BI, CA, US, so that the user can immediately identify whether a concordance line is from the British Isles, Canada, or the United States.

Corpus analysis tools enable users to investigate and manipulate the information contained within a corpus in a variety of ways. For example, most corpus analysis packages comprise a "concordancer", which will find all the occurrences of a search word, or search pattern, and display them in the centre of the screen, together with a span of co-text – a so-called Key Word In Context (KWIC) display. Figure 1 shows a KWIC display of 8 of the 70 concordance lines containing the words *discovery* or *discoveries* generated by *WordSmith Tools* from the Tourism Corpus.

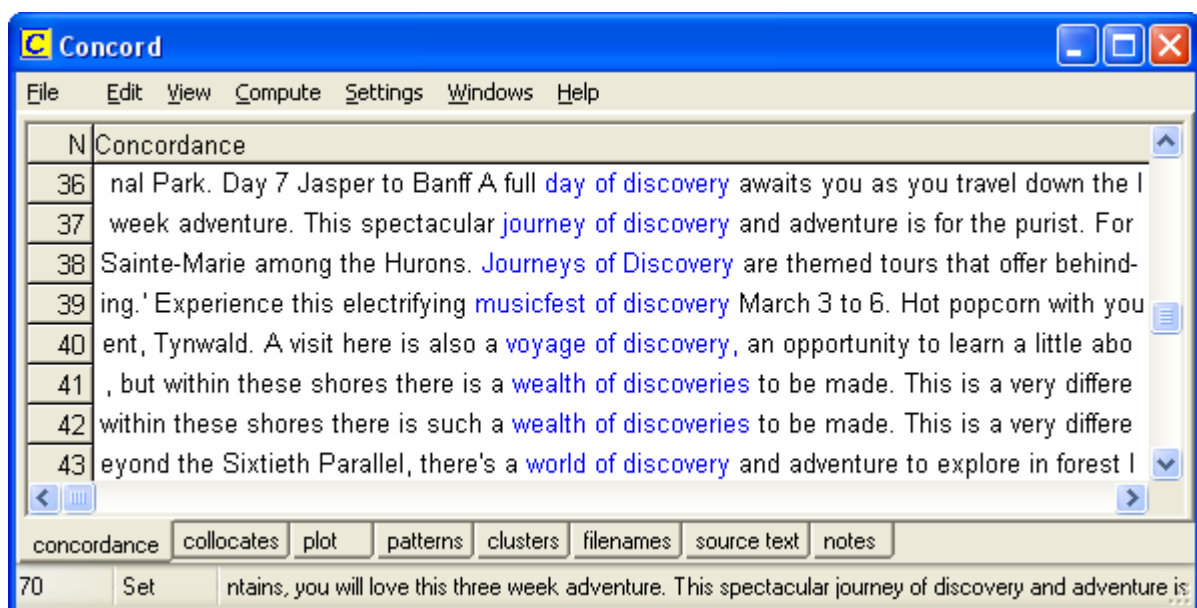


Figure 1: Some of the concordance lines generated by WordSmith Tools for the search pattern *discovery/discoveries*

You can manipulate the order of the concordance lines: for example if your search word is a noun, you can ask the concordancer to sort the words immediately preceding the search word in alphabetical order, which may help you to find suitable adjectives that collocate with the search word, as shown in figure 2.



Figure 2: Edited KWIC display generated by WordSmith Tools for the search pattern *discovery/discoveries*

By double-clicking on a line, you can view it in its full context, as in figure 3, which displays line 6 of figure 2 in a fuller context.

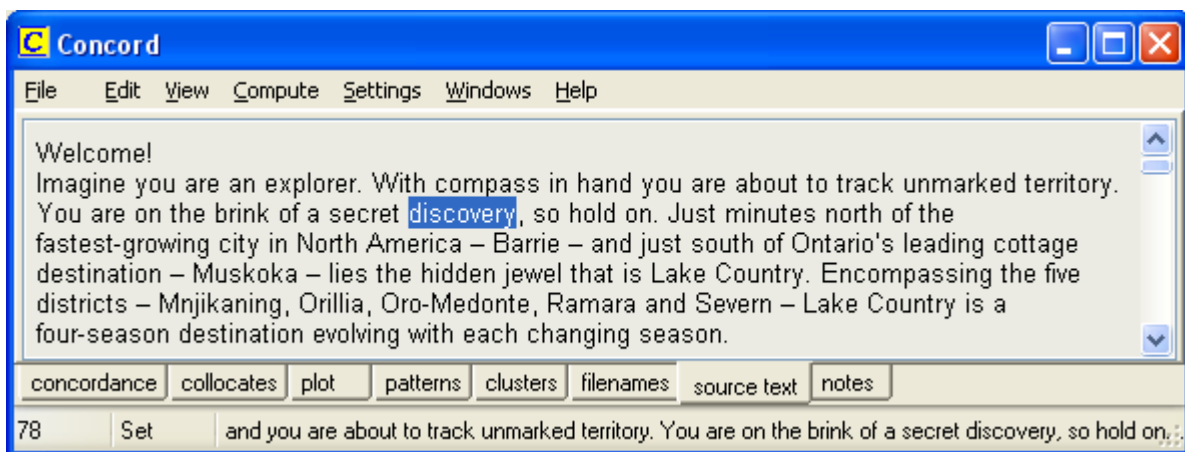


Figure 3: Display showing a concordance line in fuller context

Chance Discoveries

Bernardini (2000, 2004) is one of the leading advocates of using corpora for “discovery learning” and has encouraged advanced students of English to browse the 100-million-word British National Corpus (BNC) in open-ended, exploratory ways. In Bernardini (2001), the author describes a “journey of discovery” that she herself undertook with the BNC.

Bernardini's students have also exploited a variety of other corpora in addition to the BNC – larger and smaller, general and specific, monolingual and bilingual – and have been guided to progress from more convergent activities to autonomous browsing (Bernardini, 2002).

Zanettin (2001), describing how a relatively small corpus (250,000 words) of British newspaper articles was used as a translation aid by Italian students translating mainly from their mother tongue into English, shows that some information relevant to the translation assignment resulted from chance discoveries.

In Wilkinson (2005a), I illustrated how a specialised monolingual target-language corpus can be of great help to the translator in confirming intuitive decisions, in verifying or rejecting decisions based on other tools such as dictionaries, in obtaining information about collocates, and in reinforcing knowledge of normal target language patterns. I also touched briefly on the potential of corpora to throw up previously unknown information that may be relevant to the translation assignment at hand or may come in handy for future assignments.

The KWIC display in figure 4 illustrates some of the concordance lines generated for the search pattern *dogsle*/dog sle*/dog-sle**. The translator is looking for a translation equivalent for the Finnish term *koiravaljakkoajelu*. After hunting through traditional translation aids, the translator has come up with the terms *dog sled*, *dog sledge* & *dog sleigh*, each of which is also often written with hyphens or as one word. The corpus helps in deciding on which of these alternatives to use. The original KWIC display contained 22 hits for *dog sled*, 27 hits for *dogsled*, and 6 hits for *dog-sled*, with no hits at all for *dog sledge* or *dog sleigh* or variations thereof. Moreover there were 68 hits for *dogsledding*, often written also as two words.

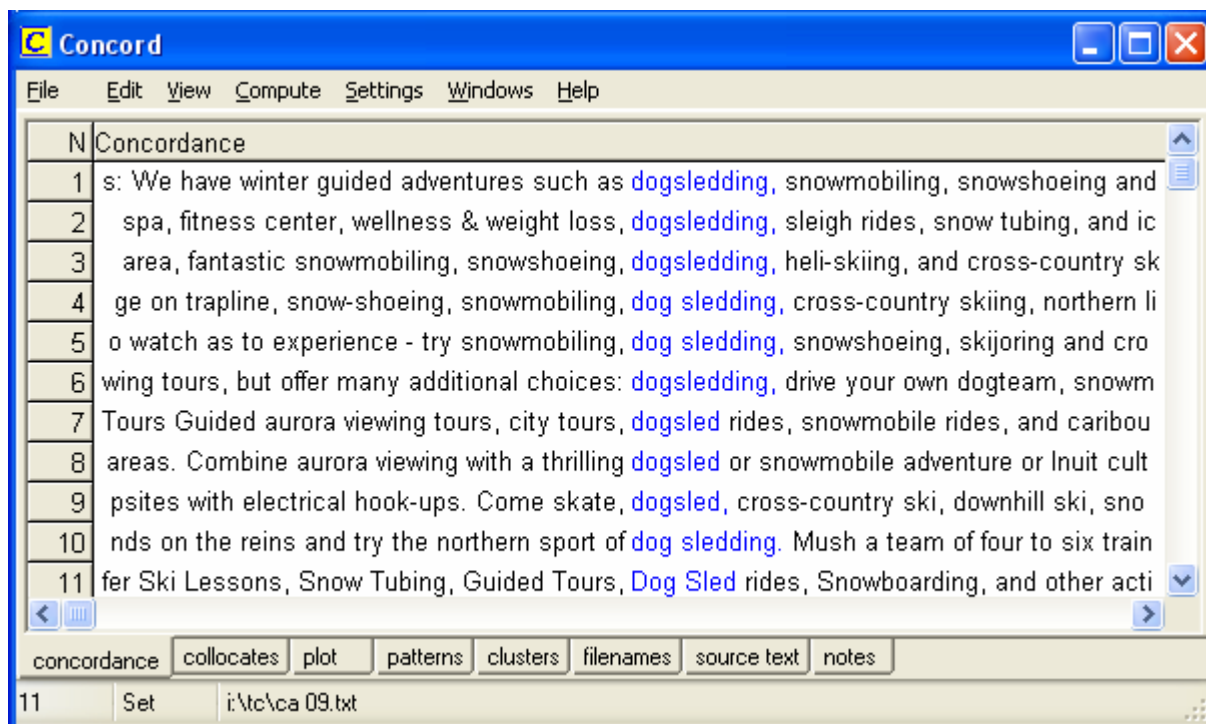


Figure 4: Edited display of some of the 118 concordance lines generated by WordSmith Tools for the search pattern *dogsle*/dog-sle*/dog sle**

But what is particularly interesting in the above KWIC display is the large amount of previously 'unknown' information the translator might acquire when browsing through it. Lines 1,3 & 4 contain references to *snowshoeing*; lines 2 & 11 mention *snow tubing*; lines 7 & 8 *aurora viewing*; line 3 mentions *heli-skiing*, line 5 *skijoring*, and line 9 *electrical hook-ups*. All of these may lead to further exploration by viewing in fuller context or by entering new search patterns.

So we can see that the search pattern *dogsle*/dog-sle*/dog sle** provides a rich source of paths to explore for those wishing to embark on a journey of discovery à la Bernardini. And indeed, in my translation courses in Savonlinna, I encourage students to explore interesting leads generated by corpus searches, record potentially-useful discoveries, and share their findings in class. Recent student-discoveries include *interpretive centre*, *float plane*, *seaplane*, *perimeter trail*, *sport fishing*, *fly-in fishing*, *fly-in resort*, *illuminated skiing loop*, *bridleway*, *skijoring* & *wildlife viewing*.

Of course, such serendipitous learning also occurs when consulting texts in printed form – when you encounter interesting 'leads' in the text, you can follow them up in other sources. However, digitalised texts allow such leads to be explored much more rapidly and systematically.

Web searches also allow for serendipitous learning. However, although the Web can be an invaluable mine of information, especially for discerning translators who have honed their search skills, it can sometimes be slow, due to the time that is often required for separating the wheat from the

chaff resulting from the numerous 'unreliable hits' that are generated. A well-designed specialised target-language corpus can probably in many cases be a more efficient and reliable tool for serendipitous learning as well as for searches that are more 'targeted'.

Advanced Searching

Unfortunately the professional translator striving to meet a deadline for a brief, or indeed the translation student trying to meet a deadline for a teacher-set assignment, often does not have the luxury of making leisurely journeys of discovery due to time pressures. Therefore it is necessary to develop other strategies for discovering "unknowns", or at least "lesser knowns" and "partially knowns". As Varantola (2002, p.180) points out, search strategies must sometimes be elaborate, and if no adequate search string or term springs to mind, translators need to think of indirect ways of finding what they are looking for.

Examples of creative searching techniques are given in Wilkinson (2005b). Such so-called "fuzzy" searches can increase the likelihood of "accidentally" finding relevant information. The *Advanced Search* feature of *WordSmith Tools* is especially useful for implementing creative searches. In my experience, newcomers to corpus analysis tools tend to under-use this feature, and therefore I shall provide a couple of examples of how it can help in discovering potential translation candidates.

The *Advanced Search* feature facilitates concordancing with contextually-relevant search words. It works in a way similar to the proximity operators used by search engines – you can restrict a concordance search by specifying a context word or context words which either must (or must not) be present within a certain number of words of your search word.

Example I: Fantastic Fishing

Suppose a Finnish translator needs to find a translation equivalent in English for the verb *pilkkiä* and/or the noun *pilkkiminen*. These occur frequently in Finnish tourist brochures. The translator knows that they mean "fishing through a hole cut in the lake ice".

A bilingual dictionary may suggest words like *jig / jigger / jigging*. If one checks, for example, *jigging* in a monolingual dictionary or on the net, one will find that this refers to the technique of jerking a jig (a small artificial lure) or other bait up and down in the water. This does not convey the fact that the activity of *pilkkiminen* takes place in winter and through the ice.

The translator might decide to go for a translation like "fishing with a jig through a hole cut in the ice." This would probably be a feasible

translation, but might be a bit long-winded if the term appears repeatedly in your text.

When translating tourism-related texts, student translators at the Savonlinna campus of Joensuu University can use *WordSmith Tools* together with the Tourism Corpus to search for translation equivalents. In this case they could enter *fishing* as their main search word, and then click on the *Advanced* tab and enter *winter* as their context word, setting the context search horizons as they see fit. In figure 5, the horizons have been set so that concordance lines will be generated whenever *winter* appears within 5 words to the left or right of *fishing*.

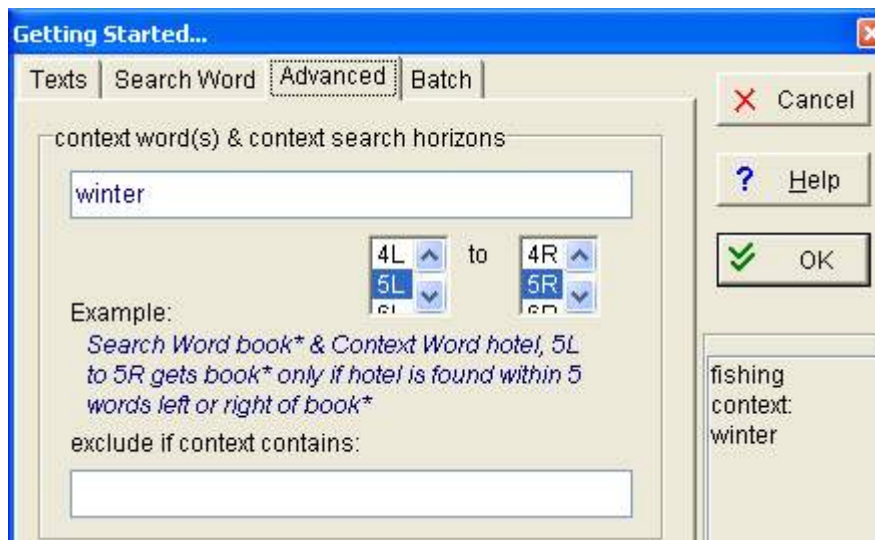


Figure 5: Search query using WordSmith Tools for *fishing* with *winter* as the context word

The resulting KWIC display is shown in figure 6. The translator will quickly notice the occurrences of the term *ice fishing* in lines 7-12 & 14-15. A follow-up search for *ice fishing/ice-fishing* without a specified context word will produce many more concordance lines, which can be explored and viewed in their full text context.

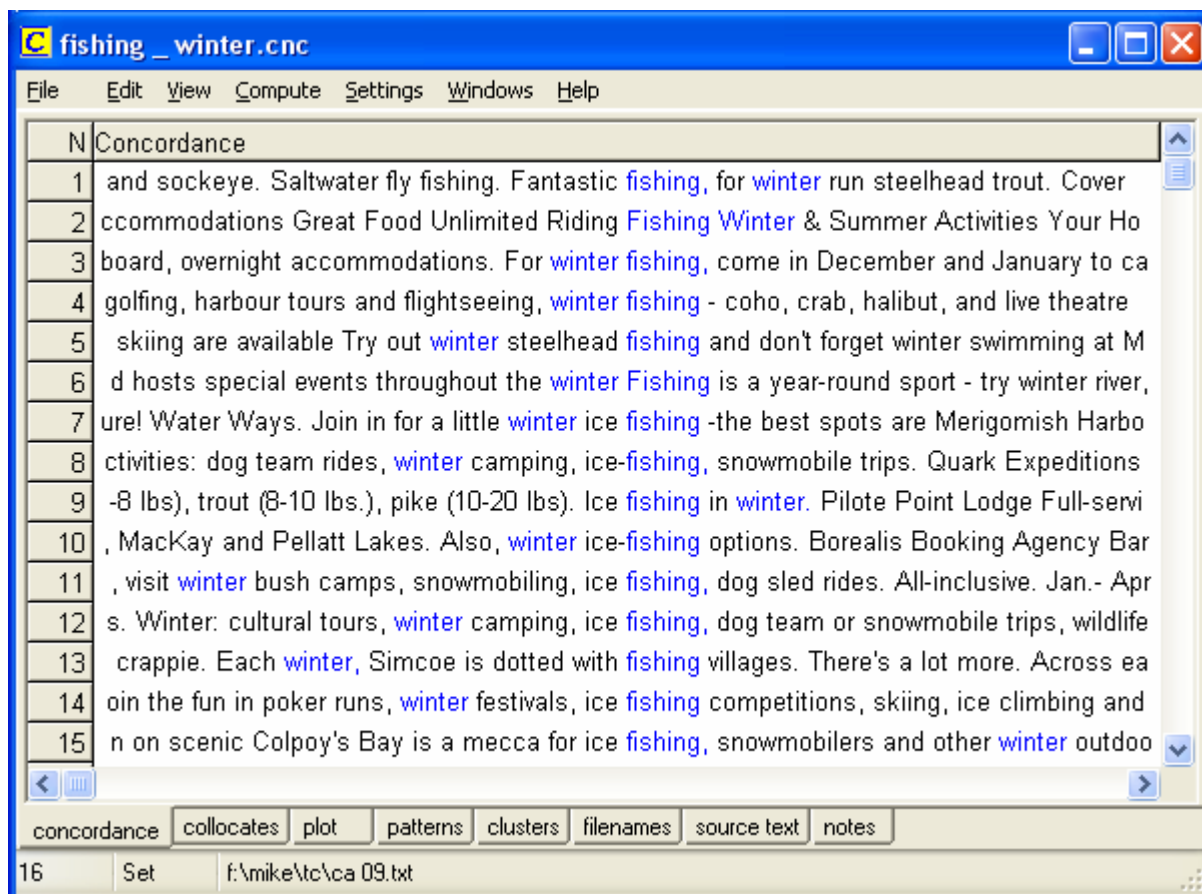


Figure 6: KWIC display for the search word *fishing* with *winter* as the context word

A follow up using other resources will quickly confirm that this is a good equivalent for the Finnish term. For example Wikipedia gives the following definition: "Ice fishing is the sport of catching fish with lines and hooks or spears through an opening in the ice on a frozen body of water. Fisherman may sit on a stool on the open expanse of a frozen lake or sit in a heated cabin on the ice with bunks and amenities". Nevertheless the translator may decide that since the term *ice fishing* appears only in Canadian brochures, target audience readers who are from other countries may not be familiar with this concept, and therefore a longer explanation may be needed the first time this term appears, after which the more concise translation of *ice fishing* can be utilised.

Example 2: Jingle Bells

Suppose our Finnish translators need to find an equivalent in English for *hevosrekiäjelu* as in the following phrase from an authentic commission:

Koiravaljakko- ja hevosrekiäjelut tilauksesta. (Dogsled rides and horsesled rides by prior booking).

