Using Keyword Tools to Explore Lexical Differences between British and American English in Specialised Corpora

Michael Wilkinson (michael.wilkinson@uef.fi)
University of Eastern Finland, Finland

Abstract
The article suggests that keyword tools have been ignored in computer-assisted language learning and computer-assisted translating, and shows how it can, for example, shed light on lexical differences between British and US tourism texts, and thus help students to make appropriate lexical choices depending on the audience they are writing for. A brief introduction to the main tools of corpus analysis programs (i.e. the concordancer, the word-list tool and the keyword tool) is followed by a survey of various reference corpora that have been utilised to generate keyword lists for corpora under investigation. For this study, two equal-size corpora – one comprising texts from British tourist brochures and the other from US tourist brochures – are used to generate keyword lists, and explanations are proposed for some of the lexical differences. Key-cluster lists as well as searches with the concordancer are also employed to supplement and aid the analysis. Though tourism texts form the basis of this study, the same approach could also be adopted to find lexical differences between different varieties of English in other specialised domains.

Keywords: American English, British English, concordancer, corpus analysis, keyword tool, lexical differences, specialised corpora

INTRODUCTION

Corpus analysis programs allow users to access, display, investigate and manipulate the information contained within an electronic text corpus in a variety of ways. Such software usually comprises an integrated suite of tools including a concordancer, a word-list tool, and a keyword tool. All three of these tools have been widely exploited in research by corpus linguists to investigate lexical, grammatical and stylistic features of texts. In addition, corpus findings have been used to develop reference materials and textbooks, including dictionaries, for language teachers and learners.

There have also been numerous case studies describing how the concordancer can be used in computer assisted language learning (CALL) or in computer-assisted translating (CAT), but very little attention has been given to the potential of the keyword tool. This article describes how the keyword tool can be exploited by language learners and is based on keyword-related assignments performed by advanced students of English at the University of Eastern Finland.

THREE CORPUS ANALYSIS TOOLS

The focus in this article is on the keyword tool, but since the word-list tool is required in order to generate keyword lists, and since the resulting lists can best be analysed by making use of the concordancer, all three of these tools are briefly described below. In this investigation the corpus analysis program WordSmith Tools version 6 (Scott, 2012) is used.

The concordancer

The concordancer typically displays all the occurrences of a search pattern in the corpus centred vertically on the screen and surrounded by their immediate co-text, as shown, for example, in Figure 1.

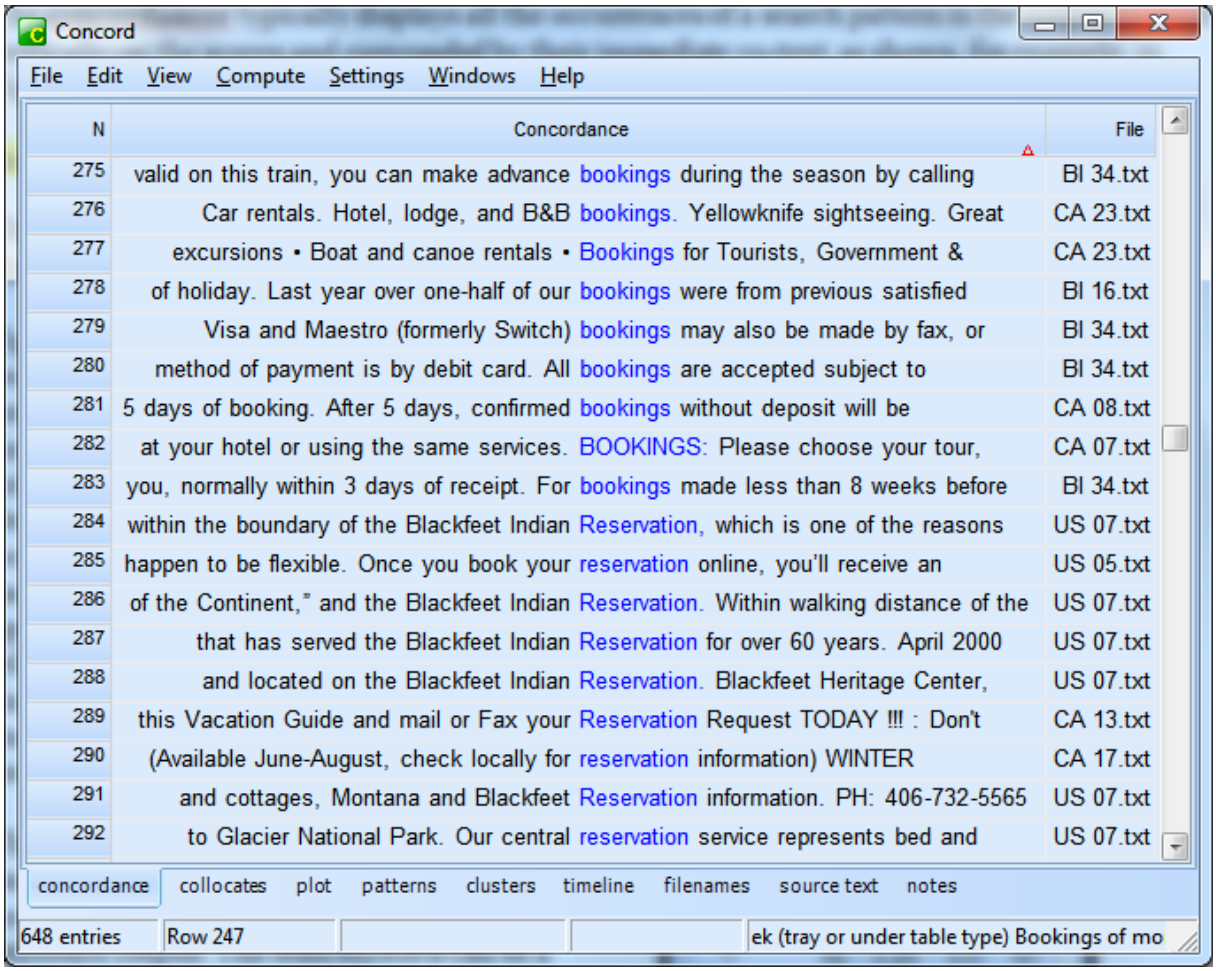


Figure 1. Concordance lines generated by a search for *booking*/reservation**.

Figure 1 shows some of the 648 concordance lines generated by a search of a one-million word corpus of texts compiled from tourist brochures for the pattern *booking*/reservation**. (The "Tourism Corpus" is described in more detail in Section 3). The results have been sorted with "Centre" as the main sort, and so lines 1-137 comprise "hits" for *booking*, lines 238-283 hits for *bookings*, lines 284-389 hits for *reservation*, and 340-648 this for *reservations*. The results could be rapidly sorted in other ways, for example with the words to the left of the search words in alphabetical order in order to study which adjectives collocate with them. The name for this kind of display – *KWIC* display (key word in context) – has become established in the literature. A better name might have been "Search Term in Context" or "Search Word in Context" in order to avoid confusion with other meanings of the term *key word*/*keyword*, such as that used in the title of this article.

Case studies that show how the concordancer can be used for CALL activities include those by Ädel (2010), Alshaar and AbuSeileek (2013), Boulton (2012), Chambers (2005), Kennedy and Miceli (2009), Mull (2013) and Varley (2009). These studies were all carried out with higher education students, whereas the case study by Braun (2007) is a rare example of integrating corpora into language learning in secondary education. For overviews of using corpora in language teaching see Gavioli (2006) and Römer (2008, 2010) and for numerous examples of the practical use of

concordancers in the classroom see Lamy and Klarskov Mortensen (2010). The concordancer can also be a very useful CAT aid, especially for students who are translating special field texts into their L2 or L3 (see e.g. Rodríguez-Inés 2013). For an overview of WordSmith's concordancer see Wilkinson (2011).

The word-list tool

The word-list tool shows all the words or word-clusters in a corpus displayed in alphabetical order or in frequency order. For example, Figure 2 shows the first 25 words in order of frequency of a wordlist that was generated from the 101 text files of the one-million word Tourism Corpus. The columns displayed here show the word, its frequency, its frequency as a percentage of all the words in the corpus, and the number of texts each word appears in. (The # symbol represents a number). The only "tourism-related" word that appears in the display is *park* in line 23, but more tourism-type words can be found further down the list. Wordlists are necessary in order to generate keyword lists, as will be explained later.

One limitation of the word-list tool when using an untagged corpus (i.e. a corpus in which words have not been assigned grammatical tags corresponding to the word class they belong to) is that the tool is unable to distinguish between homographs. The corpus used to generate the word-list shown on the right contains many examples of homographs, such as *park*, *wind* and *exhibit* (each of which can be a noun or a verb), *frequent* (which can be a verb or an adjective) as well as *content*, *present*, and *second* (each of which can be a noun, a verb, or an adjective).

N	Word	Freq.	%	Texts
1	THE	60,540	5.67	101
2	#	39,353	3.69	101
3	AND	39,058	3.66	101
4	OF	30,400	2.85	101
5	A	22,517	2.11	101
6	TO	20,158	1.89	101
7	IN	17,522	1.64	101
8	IS	10,553	0.99	101
9	FOR	10,515	0.99	101
10	WITH	8,105	0.76	101
11	ON	8,048	0.75	101
12	AT	6,984	0.65	101
13	FROM	6,114	0.57	101
14	ARE	6,027	0.56	101
15	YOU	6,021	0.56	101
16	OR	5,996	0.56	100
17	AS	4,325	0.41	101
18	BY	3,991	0.37	101
19	S	3,953	0.37	65
20	THIS	3,849	0.36	100
21	YOUR	3,812	0.36	100
22	AN	3,597	0.34	101
23	PARK	3,164	0.30	90
24	THAT	3,064	0.29	101
25	ALL	3,048	0.29	101

Figure 2. First 25 lines of a wordlist generated from a 1-million word corpus of tourism texts.

The keyword tool

The **keyword tool** shows words that occur unusually frequently (or infrequently) in the *study corpus* (SC) that is being investigated in comparison with a *reference corpus* (RC). The "keyness" of a word generated by the keyword tool is determined purely statistically – the program computes the word's frequency in the word-list of the SC, the number of words (or tokens) in the SC, its frequency in the RC, and the number of words in the RC, and cross-tabulates these. Keyness is thus not based on words that are subjectively regarded as being important (though many of the words in a keyword list will conform to expectations of importance). Hence, as mentioned earlier, *keywords* generated by the keyword tool should not be confused with the key-word-in-context feature of the concordancer, nor with any of the other meanings of keywords.

Keyword results (i.e. the ranking of the keywords) may be affected by several factors, such as the size and composition of the reference corpus (see, for example, Goh, 2011). The method of statistical analysis employed by the corpus analysis program also affects the results. In the following analyses, the keyness values have been calculated using the Log Likelihood test; WordSmith also offers the option of using the chi-square test of significance. In addition, the language settings that are used when making the word lists can have an influence – for example whether hyphens separate words, or whether apostrophes are regarded as being part of a word.

The following illustrates how the keyword tool can, for example, shed light on differences between British English (BrE) and American English (AmE) in specialised domains. Awareness of these differences can be important in helping students to make lexical choices, bearing in mind whether they are writing or translating for predominantly British, American, or multinational audiences.

THE TOURISM CORPUS

In experimenting with the keyword tool to identify differences between BrE and AmE, an untagged monolingual corpus of texts taken from tourist brochures was used. The initial version of the Tourism Corpus (hereafter referred to as the TC) was compiled by the author in 2004 to serve as an aid for Finnish students of translation, and was made accessible to staff and students for teaching and research purposes on the local network of the Savonlinna School of Translation Studies.

The intention was that the TC would be a so-called open corpus, i.e. texts would be constantly added (and some texts might be removed) to reflect the fact that language within the field of tourism marketing is, as in other special fields, constantly evolving. Consequently the TC was expanded in 2007, and at present comprises 101 text files. Texts from tourist brochures from the British Isles account for over 350,000 words, texts from Canadian brochures account for almost 360,000 words, while the US component amounts to almost 365,000 words. The expanded version of the TC was subsequently made accessible to staff and students on the local network of the Joensuu campus of the University of Eastern Finland.

The total size of the corpus amounts to around 1,075,000 words, and the corpus can be regarded as comprising three similarly-sized sub-corpora. The file names have been labelled as either BI, CA or US, so that the user can immediately identify whether a concordance line originates from the British Isles, Canada, or the United States. In the following, the British sub-corpus is referred to as the TC-BI, and the American sub-corpus as the TC-US.

Electronic corpora can be "enriched" by, for example, annotating them with part-of-speech (POS) tagging, and this is especially useful in order to enable researchers to carry out more sophisticated linguistic investigations and students to carry out more specific searches. For example, tagging would help to get round the problem of homographs mentioned in Section 2.2. However, although tagging programs have been designed to carry out such annotation automatically, checking and editing the output is time-consuming, and so the Tourism Corpus has not (yet) been tagged. Nevertheless, even an untagged corpus of texts (so-called "raw" text) can be very useful in helping students to confirm intuitive decisions, to verify or reject decisions based on other tools such as dictionaries, to obtain information about collocates (words that typically co-occur), to reinforce knowledge of normal target language patterns, and to learn how to use new expressions. For a more detailed discussion of the pros and cons of corpus annotation, see Anthony (2013, pp. 147-148).

General Reference Corpora

The keyword tool is "traditionally" used for comparing a word-list generated by the word-list tool from the corpus under investigation (usually comprised of language that is in some way specialised) with a word-list generated by the word-list tool from an "appropriate" reference corpus (usually much larger than the study corpus and often containing language of a more "general nature"). In this article the corpus under investigation is referred to as the *study corpus*, though some researchers (e.g. Scott and Tribble, 2006) refer to the corpus being investigated as the *node corpus*.

One reference corpus often used by corpus linguists for keyword analysis is the British National Corpus (BNC) – a 100 million word collection of samples of written and spoken language from a wide range of sources, designed to represent an extensive cross-section of BrE from the late 20th century. It was first released in 1995. The written part (90%) includes, for example, extracts from newspapers, specialist periodicals and journals, academic books and popular fiction, among many other kinds of text. The WordSmith Tools website contains a freely- downloadable word-list derived from the BNC. However the BNC is not perhaps an ideal reference corpus for a non-fiction corpus containing texts from several varieties of English, such as the TC, for the following reasons: (a) it consists only of BrE; (b) it contains a sizable spoken component; (c) it contains a large amount of literary texts; (d) most of the texts were published in the 1980s and early 1990s, and are therefore already somewhat dated.

Another oft-used reference corpus is the Guardian corpus, comprising news texts from the Guardian 1998-2004 and amounting to over 250 million tokens; the Guardian word-list is also freely-downloadable from the WordSmith website, and is perhaps more suitable as a reference for non-fiction corpora such as the TC, since the texts it is derived from are written, informative and relatively recent – though again it is perhaps not ideal since the texts are BrE only.

Two other corpora have often been used by corpus researchers as reference corpora, namely the Freiburg-LOB corpus of British English (FLOB) and the Freiburg-BROWN corpus of American English (FROWN). These are one-million-word corpora representing language of the early 1990s and containing texts from 15 text categories. Each of these corpora has a large literary component, which, as with the BNC, means that they are not perhaps ideal as reference corpora for non-fiction corpora. Moreover, they are unfortunately not freely available; these and other corpora have been collected together on a CD-ROM by ICAME (the International Computer Archive of Modern and Medieval English), and the cost for a single user is around €400. However, it is possible that those studying or working in higher education will have a copy that they can access at their own institutions.

Comparing the TC with a General Reference Corpus

When using the entire Tourism Corpus as the study corpus, a suitable reference corpus was created by extracting the informative components from the FLOB and FROWN corpora; these components were then combined to form a non-fiction corpus of general BrE and AmE amounting to 1.4 million words.

Figure 3 is a screenshot of the first thirty keywords of the whole TC when compared to this reference corpus. The keyness column reflects how "outstandingly frequent" the study corpus (SC) words are in comparison with those of the reference corpus (RC) – the higher the figure, the stronger the keyness. The other columns of statistics for the word *park* (ranked 2nd in order of keyness) reveal that it occurs 3,164 times in the SC, amounting to 0.30% of all the words (tokens) in the SC, whereas it occurs only 87 times in the RC – such a small figure that it does not even feature in the reference corpus percentage column.

The words that have "floated to the top" do not contain many surprises and give a fairly clear picture of the topic, or "aboutness" (Phillips, 1989), of the SC. They also indicate some stylistic features; e.g. the frequent use of *you* and *your* to address the reader in tourist brochures. Various adjectives (such as *beautiful*, *scenic*, *spectacular* and *great*) as well as proper nouns (such as *Canada* and *Vermont*) can also be found when scrolling through the top 100 keywords.

TC v General BI & US.kws

File Edit View Compute Settings Windows Help

N	Key word	Freq.	%	RC. Freq.	RC. %	Keyness
1	AND	39,058	3.66	26,685	2.22	4,186
2	PARK	3,164	0.30	87		4,091
3	LAKE	2,615	0.25	36		3,617
4	YOU	6,021	0.56	1,479	0.12	3,533
5	YOUR	3,812	0.36	480	0.04	3,364
6	WWW	1,972	0.18	0		2,980
7	S	3,953	0.37	679	0.06	2,976
8	RIVER	2,120	0.20	61		2,724
9	COM	1,731	0.16	0		2,615
10	FISHING	1,765	0.17	58		2,226
11	MOUNTAIN	1,774	0.17	61		2,222
12	ENJOY	1,587	0.15	38		2,085
13	TRAILS	1,356	0.13	5		1,989
14	TRAIL	1,403	0.13	21		1,927
15	DAY	2,363	0.22	366	0.03	1,884
16	NORTH	1,965	0.18	226	0.02	1,801
17	TOUR	1,343	0.13	50		1,661
18	HISTORIC	1,229	0.12	24		1,650
19	GOLF	1,154	0.11	11		1,633
20	WEB	1,105	0.10	4		1,621
21	TOURS	1,126	0.11	14		1,568
22	MUSEUM	1,307	0.12	67		1,524
23	AREA	2,064	0.19	367	0.03	1,521
24	OFFERS	1,294	0.12	77		1,459
25	VISIT	1,321	0.12	98		1,407
26	SKI	955	0.09	3		1,406
27	SITE	1,342	0.13	108		1,396
28	LOCATED	1,103	0.10	42		1,359
29	TOWN	1,258	0.12	89		1,358
30	MILES	1,251	0.12	93		1,332

KWs plot links clusters filenames source text notes

500 entries Row 1

Figure 3. Keyword list generated when the TC is compared with a reference corpus of general texts.

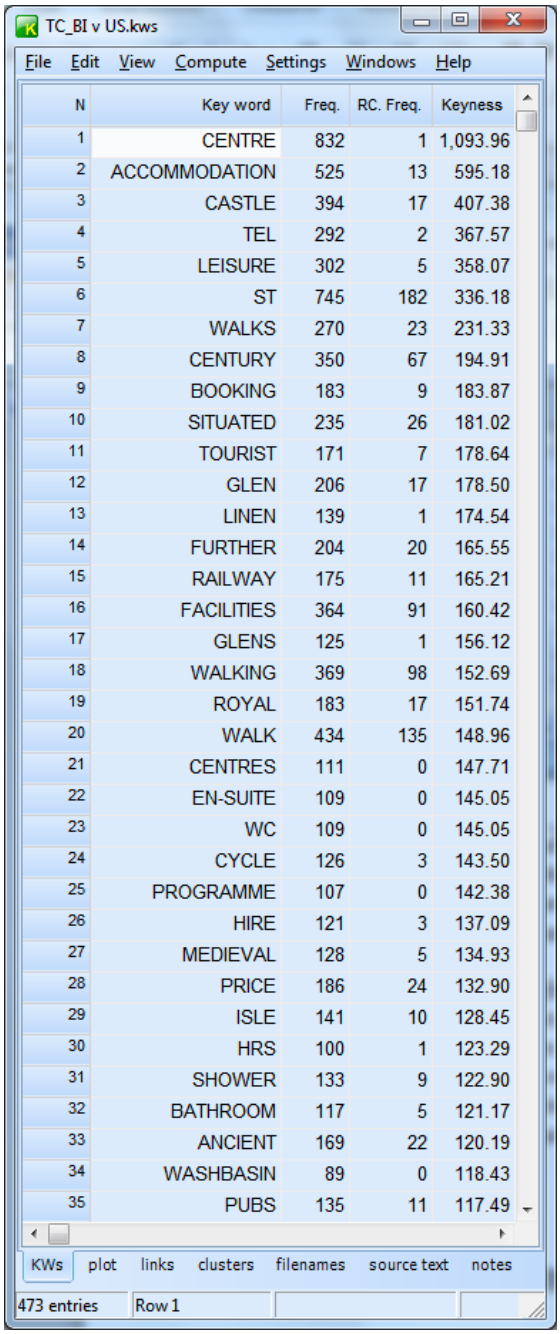
Interestingly enough, very similar results were obtained by Kang and Yu (2011). They compiled a relatively small corpus (just over 100,000 words) of Tourism English (TEC) from US and British tourism websites. This corpus was used to investigate the stylistics of Tourism English and, as part of the investigation, a keyword list was made using FLOB as a reference corpus. The words in the list bear remarkable similarity to the keywords of the TC. Words in the TEC with strong keyness include adjectives (such as *beautiful*, *spectacular*, *famous*, *grand*, *great*, *popular*, *natural*), proper nouns (such as *Roman*, *Manhattan* and *California*), "scenic nouns" (such as *lake*, *river*, *mountain*, *island*, *museum*, *bridge* and *beach*), nouns of direction (like *north*, *south* and *west*), units of measurement (such as *miles* and *acres*) and two specific verbs (*visit* and *enjoy*).

However, neither the keyword list of the TC nor the keyword list of the TEC, each of which used a corpus of general English as the reference corpus, indicate lexical differences between British and American usage in tourist brochures or tourism websites. For this, a different approach is required.

COMPARING SUB-CORPORA OF THE TC

As mentioned in Section 3.1, the usual approach for generating keyword lists is to compare the corpus being examined with a much larger reference corpus, as in the study by Kang and Yu (2011). However, to shed light on differences between the language used in British and US tourist brochures, the rather unorthodox approach has been adopted here of cross-comparing similar-sized corpora, namely sub-corpora of the TC (either the TC-BI is the study corpus and the TC-US is the reference corpus, or vice-versa). In this exploration, the focus is on content words (i.e. nouns, main verbs, adjectives and adverbs) whereas function words (i.e. auxiliary verbs, pronouns, articles, prepositions and conjunctions) are ignored.

Figure 4 shows the top 35 words of the keyword list that is obtained when the British files of the TC form the study corpus and the US files form the reference corpus. The list has been slightly edited – function words have been manually deleted, as have obvious proper nouns such as UK, Ireland, Scotland, Wales, Jersey, Belfast and Cardiff. Figure 5 shows the top 35 words of the keyword list that is obtained when the US files of the TC form the study corpus and the British files form the reference corpus. Again the list has been edited by deleting function words as well as proper nouns such as Vermont, Kentucky, Virginia, Washington, and Tahoe.


TC_BI v US.kws

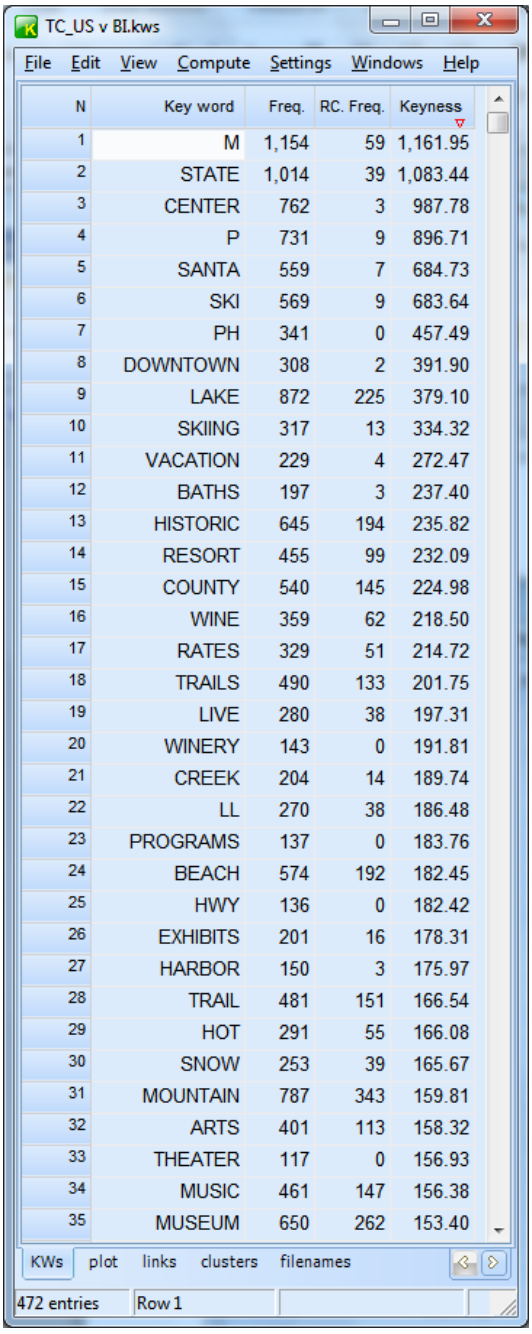
File Edit View Compute Settings Windows Help

N	Key word	Freq.	RC. Freq.	Keyness
1	CENTRE	832	1	1,093.96
2	ACCOMMODATION	525	13	595.18
3	CASTLE	394	17	407.38
4	TEL	292	2	367.57
5	LEISURE	302	5	358.07
6	ST	745	182	336.18
7	WALKS	270	23	231.33
8	CENTURY	350	67	194.91
9	BOOKING	183	9	183.87
10	SITUATED	235	26	181.02
11	TOURIST	171	7	178.64
12	GLEN	206	17	178.50
13	LINEN	139	1	174.54
14	FURTHER	204	20	165.55
15	RAILWAY	175	11	165.21
16	FACILITIES	364	91	160.42
17	GLENS	125	1	156.12
18	WALKING	369	98	152.69
19	ROYAL	183	17	151.74
20	WALK	434	135	148.96
21	CENTRES	111	0	147.71
22	EN-SUITE	109	0	145.05
23	WC	109	0	145.05
24	CYCLE	126	3	143.50
25	PROGRAMME	107	0	142.38
26	HIRE	121	3	137.09
27	MEDIEVAL	128	5	134.93
28	PRICE	186	24	132.90
29	ISLE	141	10	128.45
30	HRS	100	1	123.29
31	SHOWER	133	9	122.90
32	BATHROOM	117	5	121.17
33	ANCIENT	169	22	120.19
34	WASHBASIN	89	0	118.43
35	PUBS	135	11	117.49

KWs plot links clusters filenames source text notes

473 entries Row 1

Figure 4. First 35 keywords of TC-BI when compared with TC-US.


TC_US v BI.kws

File Edit View Compute Settings Windows Help

N	Key word	Freq.	RC. Freq.	Keyness
1	M	1,154	59	1,161.95
2	STATE	1,014	39	1,083.44
3	CENTER	762	3	987.78
4	P	731	9	896.71
5	SANTA	559	7	684.73
6	SKI	569	9	683.64
7	PH	341	0	457.49
8	DOWNTOWN	308	2	391.90
9	LAKE	872	225	379.10
10	SKIING	317	13	334.32
11	VACATION	229	4	272.47
12	BATHS	197	3	237.40
13	HISTORIC	645	194	235.82
14	RESORT	455	99	232.09
15	COUNTY	540	145	224.98
16	WINE	359	62	218.50
17	RATES	329	51	214.72
18	TRAILS	490	133	201.75
19	LIVE	280	38	197.31
20	WINERY	143	0	191.81
21	CREEK	204	14	189.74
22	LL	270	38	186.48
23	PROGRAMS	137	0	183.76
24	BEACH	574	192	182.45
25	HWY	136	0	182.42
26	EXHIBITS	201	16	178.31
27	HARBOR	150	3	175.97
28	TRAIL	481	151	166.54
29	HOT	291	55	166.08
30	SNOW	253	39	165.67
31	MOUNTAIN	787	343	159.81
32	ARTS	401	113	158.32
33	THEATER	117	0	156.93
34	MUSIC	461	147	156.38
35	MUSEUM	650	262	153.40

KWs plot links clusters filenames

472 entries Row 1

Figure 5. First 35 keywords of TC-US when compared with TC-US.

Explanations are offered below for the high keyness of some of the words that appear in the top 50 of each list, as well as for a small selection of words from lower down in the lists that appear "interesting" from a language learner's perspective. Some of the explanations are based on suggestions offered by students when performing assignments that involved comparing keyword lists of the sub-corpora of the TC as part of an online distance education course on using specialised corpora as translation aids. In the past six years, approximately 100 students of the University of Eastern Finland have completed this course.

It should be noted that it is possible to "store" several entries of a word list or a keyword list together: e.g. *ski*; *skiing*; *skis*. By grouping together word forms from the same word class under the base or uninflected form of the word, they can be analysed as a single item or "lemma". Lemmatisation can be done manually or automatically, but since the Finnish students who carried out these assignments were relatively new to corpus analysis, they were not expected to lemmatise

their corpora. For more advanced research, however, the results could perhaps be analysed more effectively if the keyword lists were based on lemmatised corpora.

Orthographic Differences

In a number of cases, the explanation for high keyness in each sub-corpus is simply due to orthographic differences between BrE and AmE. The following table illustrates this:

Table 1
Orthographic Differences between TC-BI and TC-US

	Frequency in TC-BI	Frequency in TC-US
centre(s) center(s)	943 4	1 813
programme(s) program(s)	107 3	0 261
harbour(s) harbor(s)	189 3	26 155
colour/colourful color/colorful	112 0	0 105
theatre(s) theater(s)	323 0	1 139
speciality/specialities specialty/specialties	25 2	4 110

Advanced students will probably be familiar with these differences, and if, for example, they are writing / translating for a predominantly American audience the spell-check feature of their word processor will pick up any inconsistencies in their spelling.

In addition, however, BrE and AmE sometimes differ in their usage of the singular and plural forms of certain nouns, which the spell-check won't pick up. An example of this is the word *accommodation*, which appears 525 times in the TC-BI and only 13 times in the TC-US, and thus has a very high keyness value in the TC-BI, as can be seen in Figure 4. However, further down the list of TC-US keywords, the plural form, *accommodations*, which appears 141 times in the TC-US but only 8 times in the TC-BI, can be noticed.

Culture, Climate and Geography

Certain words have high keyness in the TC-BI because of attractions related to British history and culture that are promoted in British tourist brochures. For example Figure 4 contains words such as *castle* (394 v 17), *royal* (183 v 17), *medieval* (128 v 5), and *pubs* (135 v 11). The first figure in each of the brackets shows the number of occurrences in the TC-BI, and the second figure the number of occurrences in the TC-US. Some words of Gaelic origin such as *glen(s)* (331 v 27) and *loch(s)* (83 v 0) also show high keyness. A concordance search reveals that they often occur in place names, but case sensitive searches show that when not occurring as proper nouns they are used almost exclusively when referring to Scottish and Irish scenery (see Figure 6).

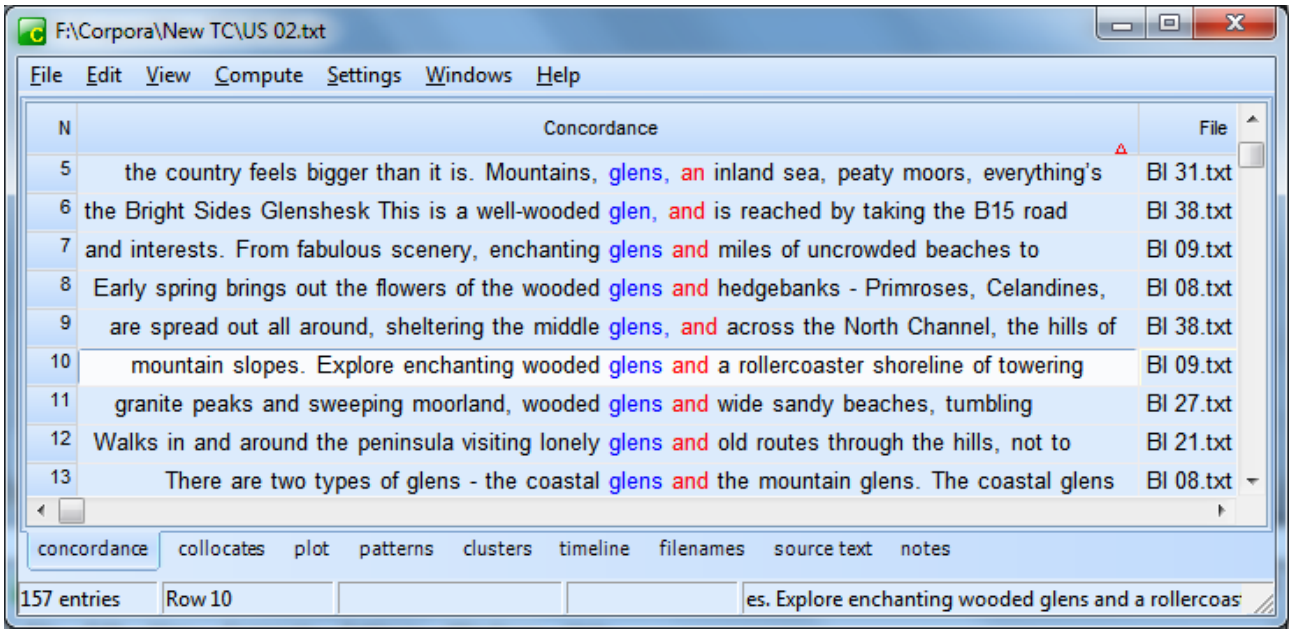


Figure 6. Concordance lines generated by a search of TC-BI and TC-US for *glen/glens.*

On the other hand certain words have high keyness in the TC-US due to the types of activities that are marketed as a consequence of geographical and climatic features. For example parts of the US have a relatively long *winter* (338 v 86) season (the first figure shows the number of occurrences in the TC-US, and the second figure the number of occurrences in the TC-BI) with plenty of *snow* (253 v 39), and it's easy to find a *mountain* (787 v 343), and so *skiers* (116 v 1) can go *skiing* (317 v 13) at *ski* (569 v 9) *resorts* (70 v 15). Unlike the US with its hundreds of ski areas, opportunities to go skiing in Britain are limited to a handful of ski resorts in Scotland. Moreover, in the US there are opportunities to go *snowmobiling* (39 v 0) on a *snowmobile* (49 v 0) or *snowshoeing* (63 v 0) on *snowshoe* (90 v 0) treks. There is a huge network of *trail/trails* (971 v 284) to cater for these various activities. Moreover, the climate in other parts of the US is much more conducive to producing *wine* (359 v 62) and so there are far more *wineries* (81 v 0) where visitors can go on *winery* (143 v 0) tours and taste various *wines* (111 v 24), though it should be mentioned that the wine industry in the UK has been steadily growing in terms of quality and stature in the past decade or so.

But why does *Santa* (559 v 7) show such high keyness in the TC-US? Do Americans promote the festive season in their tourism marketing more than the British? Or do the British usually refer to him as *Father Christmas*? And why does *hot* (291 v 55) appear in the top 30 keywords too? Is it due to the climate? Such questions can be investigated further by, for example, performing concordance searches, or by generating a keyword list of 2-word clusters, as will be illustrated below.

The differences mentioned above cast light on differences in the types of attractions and activities that tourist brochures promote on either side of the Atlantic, but they do not reveal differences in lexical usage.

Same Concept – Different Terms

Perhaps of more interest to students are same or similar concepts which are often expressed with different terms, some of which thus appear much more frequently in the TC-BI than in the TC-US or vice-versa. For example, in the TC-BI keyword list, the word *booking* appears in line 9 in Figure 4; an investigation of the keywords of the TC-US list reveals a likely equivalent – namely *reservation*. Table 2 illustrates a few more examples. A question mark (?) indicates cases where a

word has high keyness in one of the lists, but an appropriate equivalent cannot be found in the other list.

Table 2
Frequency of Near Synonyms in TC-BI and TC-US

	Frequency in TC-BI	Frequency in TC-US
autumn	47	23
fall	11	152
booking(s)	213	9
reservation(s)	16	169
?		
downtown	2	308
en-suite	109	0
?		
hire	121	3
rent/rental(s)	30	140
holiday(s)	564	225
vacation(s)	6	253
leisure	302	5
recreation / recreational	28	227
?		
RV(s)	0	61
price/prices	281	49
rate/rates	122	354
situated	235	26
located	183	466
tourist	171	7
traveler(s)/traveller(s)	42	91
tel	292	2
ph	0	341
WC	109	0
restroom(s)	0	43

As mentioned in Section 2 in conjunction with the word-list tool, it should be borne in mind that some of these words might be homographs. For example, one should have reservations about *reservation(s).* A concordance search will reveal that some of the occurrences in the corpus refer to *Indian Reservations*, as can be seen in lines 286-288 in Figure 1. Similarly, as can be seen in Figure 7, *fall* is used a number of times in the corpus in a sense that is not synonymous with *autumn*. To arrive at more accurate figures for Table 2, such considerations would need to be taken into account.

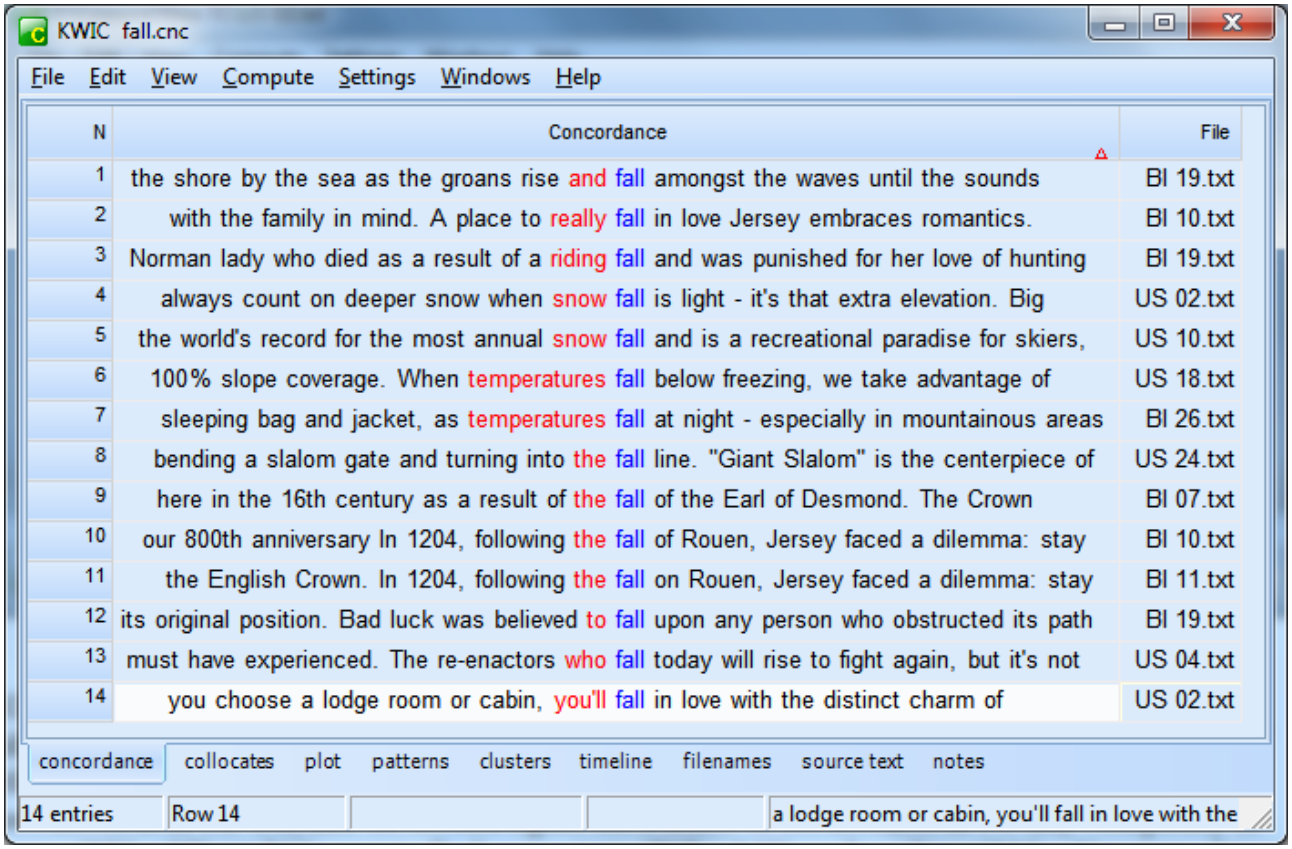


Figure 7. Edited concordance lines generated by a search of TC-BI and TC-US for *fall*.

In some of the cases in Table 2, a word has high keyness in one of the lists, but an appropriate equivalent cannot be found in the other list (as indicated with a question mark). For example the word *downtown* occurs 308 times in the TC-US, as can be seen in line 8 of Figure 5, and only 2 times in the TC-BI, but nothing similar can be found in the TC-BI keyword list. This may be because the keyword list shows only single lexical items, but the equivalent might be a compound word rather than a single word. In such cases, it can be helpful to make frequency lists of word-clusters (or "n-grams") with the word-list tool, and use these lists to generate lists of key-clusters. Figure 8 shows a keyword list generated when a list of 2-word clusters (or "bi-grams") from the TC-BI is the study corpus and a list of 2-word clusters from the TC-US is the reference corpus. As earlier, some of the obvious proper nouns have been deleted from the list.

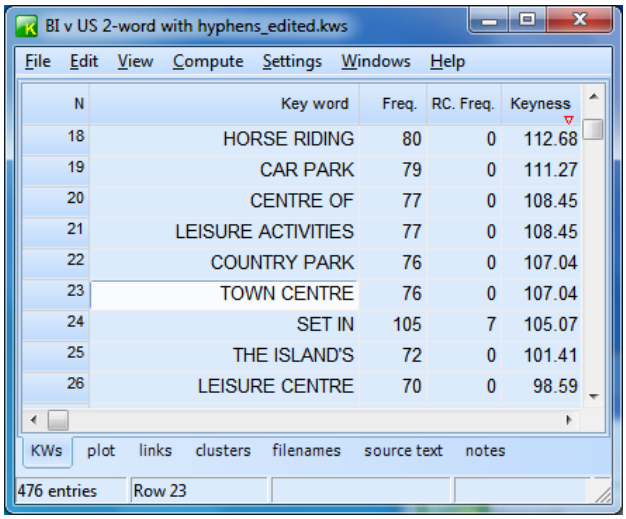


Figure 8. Two-word key-clusters of TC-BI when compared with TC-US.

Figure 8 shows that the cluster *town centre* occurs 76 times in the TC-BI and not at all in the TC-US. Even if a concordance search of the TC-US for *town center* is performed, only 10 occurrences emerge. Moreover, the two-word cluster *city centre* emerges further down the key clusters list, occurring 40 times in the TC-BI and not at all in the TC-US. Here again, a concordance search of the TC-US for *city center* produces only 4 occurrences. So it is apparent that whereas *downtown* is used exclusively in US tourist brochures, *town centre* and *city centre* predominate in British tourist brochures.

Similarly, Figure 4 shows that *en-suite* occurs 109 times in the TC-BI but not at all in the TC-US, and a follow-up concordance search for alternative spellings (*en suite/en-suite/ensuite*) reveals that there are 186 occurrences in the TC-BI and only 2 in the TC-US. No obvious synonym appears in the keyword list of single items when the TC-US is the study corpus and the TC-BI is the reference corpus. However, a 2-word key cluster list generated with the TC-US as the study corpus and the TC-BI as the reference corpus reveals high keyness for the compound word *private baths*, which occurs 51 times in the TC-US but never in the TC-BI. And a follow up concordance search for *private bath/private baths* generates 84 occurrences, only one of which is in the TC-BI. So this could be a likely US equivalent term for *en-suite*.

The keyword clusters list can also be helpful in solving some of the unexplained lexical items that appear among the top 35 items in the TC-US keyword list shown in Figure 5, such as *M* at the very top of the list and *Santa* in line 5, as well as *LL* in line 22 and *hot* in line 29.

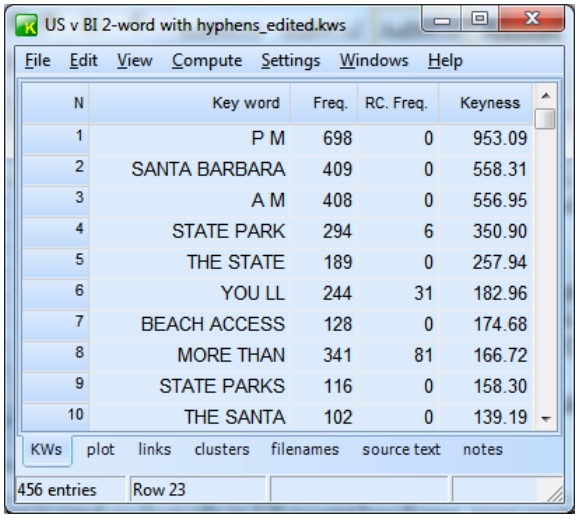


Figure 9. Two-word key-clusters of TC-US when compared with TC-BI.

In Figure 9, the TC-US is the study corpus and the TC-BI is the reference corpus. It can now be seen that *Santa* is key in the TC-US because of all the place names such as *Santa Barbara*, *Santa Cruz* and *the Santa Ynes Valley*. *LL* is key because the TC-US uses pronoun verb contractions such as *you'll* more often than the TC-BI. Further down the list there are occurrences of *hot tub* and *hot tubs*, which go part of the way towards explaining the high keyness of *hot* in Figure 5.

And Figure 9 also reveals why *P* and *M* show high keyness. It is apparent that the TC-US uses the 12-hour clock when expressing opening times, departure times and so on, in contrast to the TC-BI. A concordance search for *p.m./p m/p m* confirms this; 1428 hits in the TC-US (almost always written as *p.m.)* and only 8 hits in the TC-BI (written as *pm*). A search for *:00 will now show numerous occurrences of the 24-hour clock for expressing time in the TC-BI, but none in the TC-US (see Figure 10).

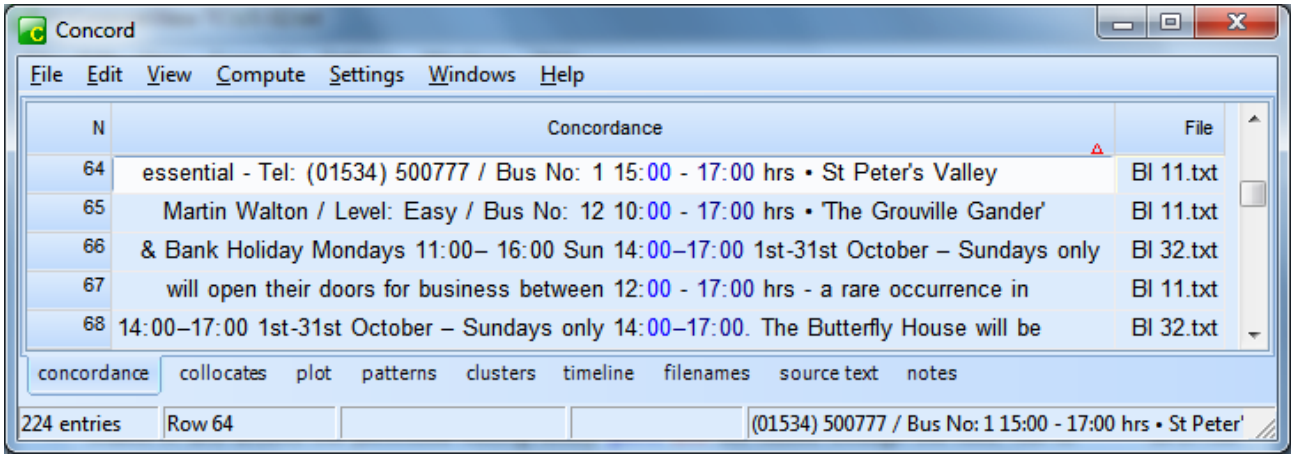


Figure 10. Concordance lines generated by a search of TC-BI and TC-US for **:00.*

The above concordance lines also show that the lexical item *hrs* (which in fact appears in the top 30 TC-BI keywords) is often used in these expressions of time. Further searches will also find occurrences of the full-stop being used rather than the colon.

As can be seen in Table 2, the lexical item *RV* (= recreational vehicle) occurs frequently in the TC-US but not at all in the TC-BI. However, it is hard to spot an equivalent in the TC-BI keyword list or in the TC-BI key clusters list. In a case like this, one solution is to try "fuzzy searches" with the concordancer (see e.g. Wilkinson, 2005). For example, a search for *motor*/motor* * will throw up words like *motor home(s) and motorhome(s)*, of which there are 5 occurrences in the TC-BI and only two in the TC-US as well as *motor caravan*, of which there are 12 occurrences in the TC-BI and none in the TC-US (see Figure 11).

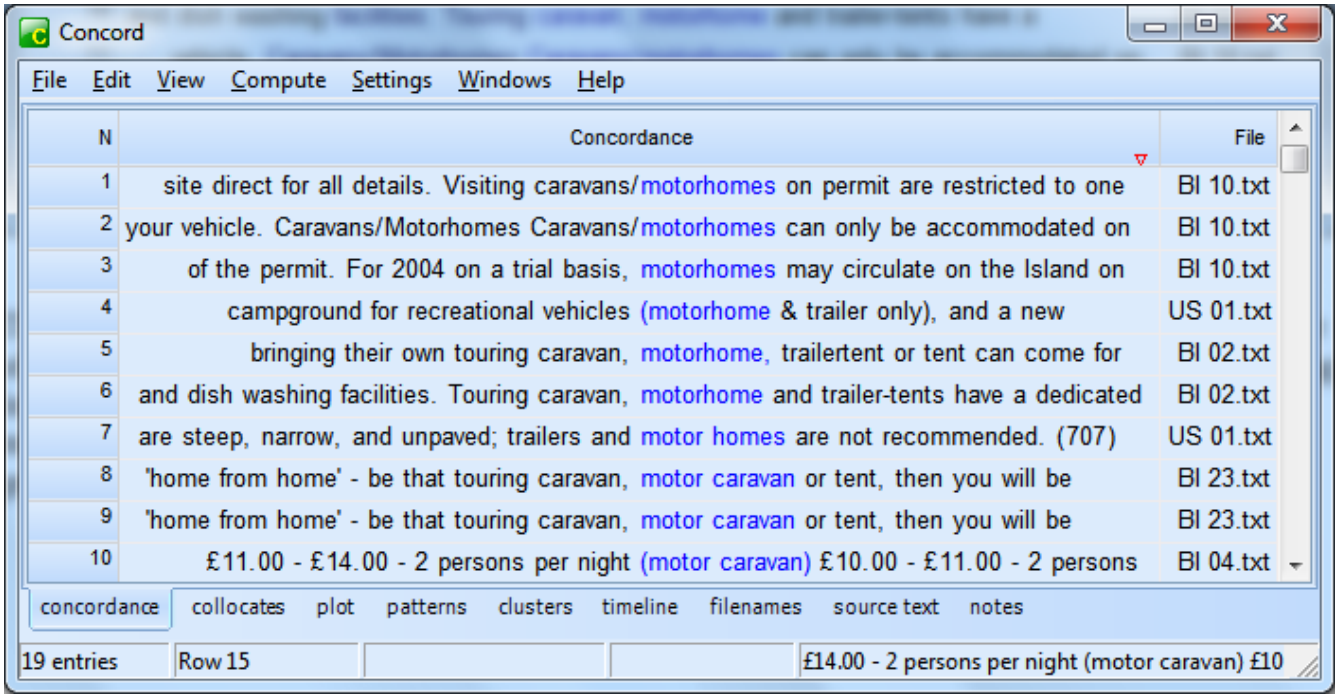


Figure 11. Concordance lines generated by a search of TC-BI and TC-US for *motor*/motor **.

These seem like possible equivalents for *RV*, but even so there are only 14 occurrences , (and none of *campervan*, which might have been expected), though one might speculate that perhaps the culture of touring in RVs is more deep-rooted in American culture, whereas touring, or ovenighting, in caravans is more common in British culture, which is confirmed by a concordance search for *caravan**, generating 70 hits, all from the TC-BI.

The examples listed in Table 2 tend to be (a) lexical items unique to one variety whose meanings are expressed by another lexical item in the other variety, such as *RV*, *WC*, *ph and en-suite* or (b) words whose meanings are actually common to both BrE and AmE but that show differences in frequency, connotation or denotation.

DISCUSSION AND CONCLUSIONS

The above analysis is not intended as an exhaustive list of lexical differences between AmE and BrE. Explanations have been offered for only a small sample of the lexical items that appear in the keyword lists – particularly in the top 50 of each list. The main aim has been to show how these lists can be used by students to find out some of the lexical differences between the TC-BI and the TC-US, or else bring out lexical items that are worth further investigation. For example, if students

peruse, say, the top 250 keywords of each list, they may come across interesting specialised vocabulary that they are not familiar with (e.g. in the TC-US: *full-service, groomed, tubing, outfitter*), and follow up their discoveries with concordance searches. (For more on serendipitous learning with corpora see Wilkinson, 2007).

In general speakers of BrE will have no problem in understanding the "American" words listed in Table 2, such as *reservation* and *rental*, just as American speakers will understand words such as *booking* and *hire*. However, when producing texts in English for a British (or European?) audience, lexical items with a strong American "flavour", such as *downtown*, *fall and ph*, should perhaps be avoided, while certain conventions, such as usage of the 24-hour clock, should be adhered to. Similarly when writing or translating for an American audience, lexical items with a strong British "flavour", such as *WC, tel. and en-suite*, should be avoided. If the text being produced is aimed at an international audience (including both Americans and the British), one should perhaps be consistent in using one variety of the language or the other, both in regard to terms and spelling.

The approach described above could be used for investigating lexical differences between other varieties of English. For example, some students at the University of Eastern Finland have used the keyword tool to explore the special features of Canadian English used in the Canadian sub-corpus of the Tourism Corpus (TC-CA). Furthermore, though this article explores lexical differences in tourist brochures, the same approach could also be used to find differences between different varieties of English in general language corpora (e.g. comparing FROWN with FLOB or with the BNC) or in corpora of other specialised domains. For example, a keyword list generated from a more highly specialised corpus consisting of, for example, British and American technical, legal, business or medical texts will probably help students to make appropriate lexical choices regarding the target audience, and will also highlight terminology that the student is not familiar with and that might warrant further investigation with the concordancer or else with dictionaries and Internet searches. Moreover, even though this article focuses on varieties of English, corpus analysis tools can handle almost any language. So the keywords approach described above could also be used to investigate lexical differences between varieties of, for example, German or Spanish.

It should also be pointed out that, although WordSmith Tools has been used in this investigation to generate keyword lists and carry out concordance searches, the same kinds of analysis can also be done with other commercially-available corpus analysis software such as MonoConc Pro (Barlow, 2004) or even freeware such as AntConc (Anthony, 2011). For more about AntConc see Wilkinson (2012), and for a more comprehensive survey of the range of software tools available for corpus analysis see Anthony (2013).

Despite the large amount of research into corpus-driven learning, corpus analysis tools have been under-exploited in the field of computer-assisted language learning (CALL). Especially in secondary schools, hands-on work with corpora is apparently still relatively rare. Similarly, corpus analysis has been under-used in the teaching of computer-assisted translation (CAT) in tertiary education. In order to accelerate the adoption of corpus analysis software in pedagogical activities, it would be necessary to "teach the teachers" by integrating corpus studies more widely into teacher education courses – see e.g. Breyer (2009) – or by arranging in-house training for tertiary-level teachers.

A major obstacle to using corpus analysis activities in the classroom is that compiling corpora is a very time-consuming process. For tips on corpus compilation see, for example, Buendía-Castro and López-Rodríguez (2013); Corpas Pastor and Seghiri Dominguez (2009); Sánchez-Gijón (2009). One solution is to involve students in the compilation process. Each student looks for appropriate texts on the Internet or in their library's online journals and converts these into plain text format and then all the texts are pooled to form a joint corpus. An example of an experiment where students

worked together in this way to build "DIY" corpora can be found in Bowker (2002). Attention should, however, be paid to the legal aspects of compiling corpora (see Wilkinson, 2006).

In their course-feedback, many students at the University of Eastern Finland have reported that they find keyword-related assignments rewarding and fun. In general, the use of corpus analysis tools, in addition to improving the quality of the "final product" (i.e. the text produced), enhances the learning experience by enabling students to be less teacher-dependent. This type of approach is often referred to as "discovery learning" or "data-driven learning" (DDL) – by interrogating and manipulating corpora and analysing data, students can make their own discoveries and deductions, and need not rely on the teacher's knowledge and intuition. In fact, through careful and critical analysis of results generated by corpus searching, combined with information obtained from other sources such as the Internet, students can often challenge and refute the teacher's suggestions. The teacher, rather than being an information provider, is more of a facilitator in the learning process, providing opportunities for students to learn through discovery and giving them hints and nudges in the right direction only when necessary.

ACKNOWLDGEMENT

Thanks to Mike Scott for permission to use screenshots of results generated by WordSmith Tools version 6. Thanks also to Mirkka Latomäki for her assistance in compiling the TC-BI and the TC-CA and to Emmi Wilkinson for her assistance in compiling the TC-US.

REFERENCES

Ädel, A. (2010).Using corpora to teach academic writing: Challenges for the direct approach. In M. Campoy-Cubillo, B. Belles-Fortuño & L. Gea-Valor (eds.). *Corpus-based Approaches to English Language Teaching* (pp. 39-55). London: Continuum.

Alshaar, A., & AbuSeileek, A. (2013). Using concordancing and word processing for improving EFL graduate students' proficiency in writing English. *The JALT CALL Journal 9*(1). Retrieved from http://journal.jaltcall.org/articles/9_1_Alshaar.pdf

Anthony, L. (2011). *AntConc* (Version 3.2.4). Tokyo, Japan: Waseda University. Available from http://www.antlab.sci.waseda.ac.jp/software.html

Anthony, L. (2013). A critical look at software tools in corpus linguistics. *Linguistic Research*, 2013, 30(2), 141-161. Retrieved from http://isli.khu.ac.kr/journal/content/data/30_2/1.pdf

Barlow, M. (2004). *MonoConc Pro* 2.2. Athelstan Publications. Available from http://www.athel. com /mp.html

Boulton, A. (2012). Beyond concordancing: Multiple affordances of corpora in university language degrees. *Languages, Cultures and Virtual Communities*. *Elsevier Procedia: Social and Behavioral Sciences*, *34*: 33-38.

Bowker, L. (2002). Working together: A collaborative approach to DIY corpora. In E. Yuste-Rodrigo (Ed.), *Language Resources for Translation Work and Research*, LREC 2002 Workshop Proceedings, 29–32., Las Palmas de Gran Canaria, Spain, 29-32. Retrieved from http://www. lrec-conf.org/proceedings/lrec2002/pdf/ws8.pdf

Braun, S. (2007). Integrating corpus work into secondary education: from data-driven learning to needs-driven corpora. *ReCALL 19*(3), 307-328.

Breyer, Y. (2009). Learning and teaching with corpora: reflections by student teachers. *Computer Assisted Language Learning 22*(2), 153-172.

Buendía-Castro, M., & López-Rodríguez C.I. (2013). The Web for corpus and the Web as corpus in translator training. *New Voices in Translation Studies*, *10*, 54-71. Retrieved from http://www.

iatis.org/images/stories/publications/new-voices/Issue10-2013/articles/article-buendia 2013b.pdf

Chambers, A. (2005). Integrating corpus consultation in language studies. *Language Learning and Technology*, 9, 111–125. Retrieved from http://llt.msu.edu/vol9num2/chambers/default.html

Corpas Pastor, G., & Seghiri Dominguez, M. (2009). Virtual corpora as documentation resources: Translating travel insurance documents (English-Spanish). In A. Beeby, P. Rodríguez Inés & P. Sánchez-Gijón (Eds.), *Corpus Use and Translating: Corpus use for learning to translate and learning corpus use to translate* (pp. 75-107). Amsterdam: John Benjamins.

Gavioli, L. (2006). *Exploring Corpora for ESP Learning.* Amsterdam: John Benjamins.

Goh, G-Y. (2011). Choosing a reference corpus for keyword calculation. *Linguistic Research 28*(1), 239-256. Retrieved from http://isli.khu.ac.kr/journal/content/data/28_1/13.pdf

Kang, N., & Yu Q. (2011). Corpus-based stylistic analysis of tourism English. *Journal of Language Teaching and Research 2*(1), 129-136. Retrieved from http://ojs.academypublisher.com/index.php/jltr/article/view/0201129136/2494

Kennedy, C., & Miceli, T. (2010). Corpus-assisted creative writing: Introducing intermediate Italian learners to a corpus as a reference resource. *Language Learning and Technology, 14*(1): 28–44. Retrieved from http://llt.msu.edu/vol14num1/kennedymiceli.pdf

Lamy, M-N., & Klarskov Mortensen, H. J. (2011) Using concordance programs in the modern foreign anguages classroom. In G. Davies (ed.), *Information and Communications Technology for Language Teachers (ICT4LT)*, Module 2.4. Slough, Thames Valley University. Retrieved from http://www.ict4lt.org/en/en_mod2-4.htm

Mull, J. (2013). The learner as researcher: Student concordancing and error correction. *Studies in Self-Access Learning Journal, 4(*1), 43-55. Retrieved from http://sisaljournal.org/archives/mar13/mull

Phillips, M. (1989). Lexical structure of text. *Discourse Analysis Monographs* 12, Birmingham: University of Birmingham.

Rodríguez-Inés, P. (2013). Electronic target-language specialised corpora in translator education: building and searching strategies. *Babel 59*(1): 57-75.

Römer, U. (2008). Corpora and language teaching. In A. Lüdeling & M. Kytö (eds.), *Corpus Linguistics: an International Handbook, Volume 1* (pp.112-131). Berlin: Mouton De Gruyter. Retrieved from http://www.lexically.net/wordsmith/corpus_linguistics_links/Roemer%20208%20HSK%20CL%20chapter%20final%20print%20version.pdf

Römer, U. (2010). Using general and specialized corpora in English language teaching: Past, present and future. In M. Campoy-Cubillo, B. Belles-Fortuño & L. Gea-Valor (eds.), *Corpus-based Approaches to English Language Teaching* (pp.18-35). London: Continuum.

Sánchez-Gijón, P. (2009). Developing documentation skills to build do-it-yourself corpora in the specialised translation course. In A. Beeby, P. Rodríguez Inés & P. Sánchez-Gijón (Eds.), *Corpus Use and Translating: Corpus use for learning to translate and learning corpus use to translate* (pp. 109-127). Amsterdam: John Benjamins.

Scott, M. (2012). *WordSmith Tools* version 6, Liverpool: Lexical Analysis Software. Available from http://www.lexically.net/wordsmith/version6/index.html

Scott, M., & Tribble, C. (2006). *Textual patterns: keyword and corpus analysis in language education*. Amsterdam: John Benjamins.

Seghiri Domínguez, M. (2008). Creating virtual corpora step by step. In P. Sánchez Hernández, P., Pérez-Paredes, P., Aguado Jiménez , & R. Criado Sánchez (Eds.), *Researching and Teaching specialized languages: New contexts, new challenges* (pp. 435-449).

Varley, S. (2009). I'll just look that up in the concordancer: Integrating corpus consultation into the language learning environment. *Computer Assisted Language Learning, 22*(2), 133–152.

Wilkinson, M. (2005). Discovering translation equivalents in a tourism corpus by means of fuzzy searching. *Translation Journal*, *9*(4). Retrieved from http://translationjournal.net/journal/34corpus.htm

Wilkinson, M. (2006). Legal aspects of compiling corpora to be used as translation resources. *Translation Journal*, *10*(2). Retrieved from http://translationjournal.net/journal/36corpus.htm

Wilkinson, M. (2007). Corpora, serendipity & advanced search techniques. *The Journal of Specialised Translation*, 07. Retrieved from http://www.jostrans.org/issue07/art_wilkinson.php

Wilkinson, M. (2011). WordSmith Tools: The best corpus analysis program for translators? *Translation Journal*, *15*(3). Retrieved from http://www.bokorlang.com/journal/57corpus.htm

Wilkinson, M. (2012). The best freeware corpus analysis program for translators? *Translation Journal*, *16*(2). Retrieved from http://www.bokorlang.com/journal/60corpus.htm