Meta
Journal des traducteurs
Translators' Journal

META

Is Bigger Better? Corpus and Dictionary Use in the Search for Compounds, Collocations, Derived Forms and Fixed Expressions

Phaedra Royle, Isabelle Richardson, Sophie Boisvert et Nicolas Bourguignon

Volume 54, numéro 3, septembre 2009

URI : https://id.erudit.org/iderudit/038312ar
DOI : https://doi.org/10.7202/038312ar

Aller au sommaire du numéro

Éditeur(s)

Les Presses de l'Université de Montréal

ISSN

0026-0452 (imprimé)
1492-1421 (numérique)

Découvrir la revue

Citer cet article

Royle, P., Richardson, I., Boisvert, S. & Bourguignon, N. (2009). Is Bigger Better? Corpus and Dictionary Use in the Search for Compounds, Collocations, Derived Forms and Fixed Expressions. *Meta*, *54*(3), 520–532. https://doi.org/10.7202/038312ar

Résumé de l'article

La création d'entrées dans le cadre de l'élaboration d'un dictionnaire bilingue mobilise habituellement des dictionnaires unilingues dans les langues source et cible, des dictionnaires bilingues ainsi que des corpus textuels. En outre, la fréquence élevée de certains mots dans le corpus impose une sélection des collocations, des mots composés, des formes dérivées et des expressions figées à inclure dans le dictionnaire. Le présent article offre un aperçu des avantages découlant de la combinaison de l'usage des sources dictionnairiques et des corpus de données. Nous proposons que la recherche par fréquence est un paramètre particulièrement utile pour résoudre les difficultés posées par l'étude de mots présents dans le corpus à une fréquence élevée.

Is Bigger Better? Corpus and Dictionary Use in the Search for Compounds, Collocations, Derived Forms and Fixed Expressions

PHAEDRA ROYLE
Université de Montréal, Montréal, Canada
phaedra.royle@umontreal.ca

ISABELLE RICHARDSON
Translator, Montréal, Canada
isabelletrad@yahoo.ca

SOPHIE BOISVERT
Translator, Montréal, Canada

NICOLAS BOURGUIGNON
Université de Montréal, Montréal, Canada
nicolas.bourguignon@umontreal.ca

RÉSUMÉ

La création d'entrées dans le cadre de l'élaboration d'un dictionnaire bilingue mobilise habituellement des dictionnaires unilingues dans les langues source et cible, des dictionnaires bilingues ainsi que des corpus textuels. En outre, la fréquence élevée de certains mots dans le corpus impose une sélection des collocations, des mots composés, des formes dérivées et des expressions figées à inclure dans le dictionnaire. Le présent article offre un aperçu des avantages découlant de la combinaison de l'usage des sources dictionnairiques et des corpus de données. Nous proposons que la recherche par fréquence est un paramètre particulièrement utile pour résoudre les difficultés posées par l'étude de mots présents dans le corpus à une fréquence élevée.

ABSTRACT

In the course of the development of a bilingual dictionary, a number of monolingual source language and target language dictionaries, bilingual dictionaries, and text corpora are typically used as tools to create entries. When dealing with words that occur at a high frequency in the corpus, determining which collocations, compounds, derived forms and fixed expressions are to be included in the dictionary is an additional complication. This paper presents the relative merits of using dictionary and corpus sources for searching for this type of information. We present frequency searching as an efficient and useful tool for corpus analysis, especially in the case of high-frequency words.

MOTS-CLÉS/KEYWORDS

analyse de corpus, recherche par fréquence, mots composés, collocations, dérivations, expressions figées
corpora, frequency search, compound, collocation, derivation, fixed expression

> […] there are really good reasons for building corpora, and as far as I am concerned, the bigger the better. (Fillmore 1991)

1. Background

During the process of lexical entry creation for a bilingual dictionary, one makes use of a number of different types of source documents in order to ensure that the entries have complete phonological, orthographic, morphological, syntactic and semantic information. A traditional reference source is other dictionaries (mono- and bilingual) in the same languages. Another source, now more and more widely used, is a lexicographic corpus – a collection of texts that have been entered into a text retrieval software and that can be consulted in order to confirm or insert additional semantic, syntactic, morphological, and orthographic information about usage for a given headword (Filmore 1992; Sinclair 1991). Provided they are, to some extent, "designed for use in the creation of dictionaries" (Atkins and Rundell 2008), corpora can prove useful not only for the identification of neologisms, but also erosion in the use of certain word meanings. They can also yield informative material on extended usage (morphological processes such as conversion or derivation), syntactic structures and orthographic changes that have not yet been attested in more traditional dictionary sources.

When consulting the corpus, a pre-determined amount of data is typically printed out, 100 contexts in the case of our dictionary (see method below). These linguistic environments are usually sufficient to provide the required contextual characterization of the word use in context, even if it has a number of different meanings and specific syntactic structures in which it is found. However, when a word has a very high frequency (such as *big*), a list of only 100 contextual examples will probably be too shallow compared to the various uses this form offers. In addition, if this word is used with multiple meanings or in a large number of compounds, collocations and fixed expressions, the data will probably not be sufficient to provide a representative sample of its various meanings and structures. What strategy should then be adopted in order to acquire a comprehensive overview of its linguistic usage? One approach is to carefully analyze the dictionary sources for clues as to the potentially interesting structures associated with the headword in question, then check the corpus to find whether these specific structures actually exist, retaining those confirmed through the corpus analysis. However, the question remains: Have we missed anything? How do we know that we have provided a full account of the adjective *big* if we were not able, due to the sheer size of the corpus, to analyze all possible structures associated which our headword? We might have overlooked something quite significant that was not included in the dictionary sources. In this respect, printing out a larger data sample will hardly be of any help. For example, our corpus search generates 110 146 hits for *big*. It is improbable that simply doubling the number of contexts will give us a representative sample of the corpus, as this would give us a sample representing less than one percent of the corpus (0.0018, to be exact). In fact, we would have to extract at least one thousand contexts from the corpus. This inefficient overabundance of data poses a series of additional problems.

There is, however, a solution to these problems. Text retrieval software often includes frequency search functions that allow for the extraction of the most frequent forms or combinations of forms in a text database. In this article we present the use

of frequency searching as a tool for the analysis of large corpora, with the specific goal of extracting information about compounds, collocations, derived forms and fixed expressions, which will subsequently be used in the *Bilingual Canadian Dictionary* (*BCD*) under the headword *big*.

1.1. The Bilingual Canadian Dictionary

The *Bilingual Canadian Dictionary (BCD)* (Roberts, Auger *et al.* 1998) is an inter-university project, funded by the Social Sciences and Humanities Research Council of Canada, whose goal is to provide a translation tool for sophisticated language users of Canadian varieties of French and English. No bilingual dictionary has been published in Canada since the *Dictionnaire canadien* by Jean-Paul Vinay and collaborators (1962). A number of English-French bilingual European dictionaries, such as *Robert-Collin*s, *Oxford-Hachett*e, and *Larouss*e, do exist, but the need for a Canadian bilingual dictionary remains (Roberts and Clas 1999). The main reason for this is that Canadian varieties of French and English are different from their European (and American) counterparts. Words can have different meanings, spellings and usage in Canada from those traditionally found in European translation dictionaries. Three universities have thus collaborated in the development of the *BCD* (the University of Ottawa, the University of Montreal and Laval University, in Quebec City) in order to provide the Canadian community with a translation tool that is adapted to their needs and available to the uninitiated public as well as specialised users.

According to Roberts and Clas (1999), the main objective of the *BCD* project is to "produce a bilingual Canadian dictionary that is a reflection of French and English as they are used in Canada." Thus canadianisms (specific words, or words that have specific meanings, orthographies, collocations or fixed expressions in Canada) as well as words and structures with international usage will be included in this information source. This dictionary has the particularity of containing a large number of compounds, collocations and fixed expressions with the express goal of aiding professionals working with both languages.

2. Method

The development of the *BCD* has had an impact on research methods geared toward the creation of a bilingual lexicography. One important feature of this project is the relative importance given to the corpus as a source for sense divisions, compounds, collocations and fixed expressions (Roberts and Clas 1999). A database (TEXTUM) was expressly created for the project by using various written text sources from Canada, the United States and France (see Roberts and Clas 1999, for details), and was hosted at the University of Montreal. English corpora include journalistic texts from Canadian and American dailies (e.g., *The Gazette, The Wall Street Journal*), excerpts from magazines, novels, scientific books and articles from Canadian editors (e.g., *Queen's, Canadian Geographic*) and scientific reports from the *American Department of Energy*. The combined size of these corpora is over 2000 million words. French corpora include journalistic texts from Canadian French dailies, weeklies and monthlies (e.g., *La Presse, Voir, L'Actualité*), novels (*Leméac*) and scientific vehicles (*ACFAS*), as well as two hexagonal French newspapers (*Le Monde* and *Ouest-France*)

totalling over 1000 million words. This allowed us to analyze word use in a variety of contexts and to uncover meaning distinctions as well as structural variations. Hapax legomena are thus relatively rare in this corpus. The texts were mostly journalistic, but also contained literary and scientific, as well as more popular styles, and allowed for comparisons between English and French usage. In addition, these texts were written for a larger public and thus are assumed to contain accepted and common language usage. TEXTUM was divided into eleven different corpora (e.g., *English Canadian Press,* or *ECP; The Wall Street Journal* or *WSJ; Queens,* etc.) and could thus be consulted one by one. This allowed for usage verification in different countries and types of texts. These texts were not lemmatised. The corpora were searched using PAT 3.3 (Fawcett 1988), a text searching application.

These corpora were used throughout the lexical entry creation process. When developing an entry, a lexicographer made use of approximately 100 key words in context (KWICS, see Atkins and Rundell 2007: 104-105) from the corpora, in addition to a number of unilingual and bilingual dictionary sources (see Appendix 1 for a list of dictionaries used). The KWICS were first used to verify if different meanings for a given word (sense divisions) proposed by the unilingual source language dictionaries were found in the corpus, and in what proportion. At this point, the lexicographer could also decide, if competing orthographies existed, which spelling to use for the word. More importantly for the purpose of this article, the lexicographer also decided what types of morphological and syntactic structures (compounds, expressions, derived words, …) the word was found in, and detected the types of collocations associated with it. The full process of documentary research for English-French translation is described below.

In order to prepare a dictionary entry, the lexicographer had in her possession a file containing paper documentation (photocopies from entries in bilingual and unilingual dictionaries) in addition to a corpus search for the headword that was to be translated. A series of KWICS were printed depending on the following factors: If the word was polysemous, 100 KWICS were printed out from the English data bases. Sixty of these were taken from the *ECP* and 20 each from *Queen's* and *WSJ*. In the case where the word was polysemous and the *ECP* database provided only a small number of contexts for it, all occurrences of the word would be printed out from all three above-mentioned English data bases. These KWICS were randomly extracted by the text searching system (PAT). If different orthographies existed, the lexicographer was responsible for verifying the relative frequency of occurrence of each one thus using the most frequent one for the headword in the dictionary.

Following this step, the lexicographer focused on the question of relative frequency of different meaning use for a headword. Unilingual English dictionaries were used for the purpose of identifying polysemy. Second, the lexicographer was responsible for confirming these multiple senses by verifying usage in the corpus. The corpus can occasionally reveal new (or old) usages of the headword that are not found in the unilingual dictionaries. This can be especially useful for the identification of new meanings and canadianisms. Finally, usage frequency in the corpus helped define the structure of the headword in the dictionary as meanings were presented in order from most to least frequent, based on corpus data.

In addition, a syntactic analysis of a lexeme's grammatical category had to be made throughout the process of corpus analysis, since TEXTUM was not lemmatised.

This was especially important when two syntactic categories were homophonous (e.g., *big* adjective or adverb, e.g., *think big*).

In the *BCD*, collocations, compounds and fixed expressions were defined as follows. Collocations are understood to be common word combinations that are not syntactically constrained (Firth 1968; Sinclair 1991; Smadja 1993; Moon 2008). Some word dyads or triads "go together" almost automatically, whereas they resist combination with other words. Failure to observe these constraints results in a lexical violation. For instance, one does not say **bleed greatly* but *bleed profusely* in English. Collocations were not considered "fixed expressions" or "phrasemes" since their meaning could be transparently understood from the meaning of each of the elements of the word combination (Clas 1994).

Compounds were all complex words (i.e., multi-morphemic units that function like simple lexical units) containing the headword, whether their meaning of *big* was semantically transparent or not (e.g., the exocentric compound *bighorn* is a type of sheep, not a **big horn*). Compounds could be written in one of three ways: as a single orthographic word (e.g., *heartbeat*), as a hyphenated orthographic word (e.g., *roller-skate*) or as two independent orthographic words (e.g., *patch test*). Two important tests serve to distinguish collocations from compounds: stress pattern and the polarity test. In English, compounds – whether written as one word, hyphenated or two words – bear only one main stress (e.g., *blackbird*) while collocations are composed of (at least) two independent lexemes and have two main stresses (e.g., *black bird*). Similarly, collocations do not allow for the insertion of a contradicting item without resulting in an inconsistency, whereas compounds do (compare *a white blackboard* vs **a white black board*). Compounds were listed alphabetically under the headword of the first component of the complex word (if it was the first word that carried meaning) when they were written as two words or hyphenated, and listed as separate entries when they were written as one word (see examples in Appendix III below).

Fixed expressions included (a) exocentric units which were not compounds (whose meaning cannot be deduced from the sum of its parts and resists modification with synonyms, e.g., *what's the big/*large deal*, [Cowie 1994]), (b) clichés and sayings, and (c) commonly used proverbs. Examples of these are (a) *To beat around the bush*, (b) *Nature abhors a vacuum* and (c) *An apple a day keeps the doctor away*. A combination of words can be considered fixed (a) if one cannot understand the meaning of the whole by looking at the meaning of the parts, and (b) if one cannot add or substitute words within it (e.g., you cannot say *A green apple a day keeps the doctor away*). As in the case of compounds, fixed expressions were listed alphabetically under the headword of the first lexical word of the expression.

2.1. The case of the adjective big

A first analysis of entries for *big* in unilingual English dictionaries was performed in order to identify its different meanings (or *sense indications*). During this preliminary analysis, we noted all possible collocations, compounds, and fixed expressions associated with *big*, as well as derived forms in its morphological family. The analysis of dictionary sources resulted in a list of possible collocations, compounds and fixed expressions to be included in the *big* family, either as separate entries or under the headword *big*.

A second analysis was performed by searching the corpus for *big*. Again, different meanings, derived forms, collocations, compounds and fixed expressions were extracted from these sources. Since 100 KWICS offered us an extremely limited sample of grammatical structures into which *big* could be inserted, we used a special function in our data base search program, termed "frequency searching," that allowed us to quickly get an overview of possible collocations, compounds, derived words and fixed expressions in the corpus. For example, if we wanted to find out the most frequent compounds or derived words starting with *big* in a given corpus, we typed the following command after having selected the corpus: [signif "big"]. If we wanted to find the most common words or phrases following[1] *big* in the same corpus, we typed the following command (with a space after *big*): [signif "big"]. Examples of results for these searches (cleaned) in WSJ and ECP are presented in Appendix II.

In the following section, we will present a review of our searches using dictionary sources and corpora, and compare results found using both of these search methods. We will outline advantages and disadvantages found when using either of these sources for derived forms, compounds, collocations and fixed expressions.

2.2. Dictionary sources

Dictionary sources listed a large number of compounds beginning with *big*. A lesser number of derived forms, collocations and fixed expressions was also listed. A number of derived and compound forms, collocations, and expressions were found in most of the dictionary sources. Examples of these are: *biggish* (derived form), *Big Apple* (compound), *big sister* (collocation, with the meaning older sister, *big* here being used in its 'old' sense), and *to be too big for one's boots/breeches/britches* (fixed expression). However, a large number of derived forms, compounds, collocations, and fixed expressions are found only in one or a small number of dictionary sources. Examples of these are: *bigly* (derived form), *big boys* (compound, meaning one having status or power), *the bigger the better* (collocation), *to be no big deal* (fixed expression, meaning not important). The lexicographer was to decide on which of these different items to include in the entry under the headword *big*. One could have used the rule of thumb that items listed in a large number of dictionaries, like *big business* (compound), should be included, based on a high level of agreement between the different sources. However, some items were present in many dictionaries without being very present in everyday usage, at least not in the corpora we used for our searches (e.g., *big noise*). We therefore used the corpus to verify if the items proposed by the dictionaries were representative of actual language use. In addition, the dictionaries did not always provide a large number of collocations or fixed expressions that could be included in our entry (and some, like the *Oxford Reference Dictionary,* 1986, provided very little information at all about *big*), and these items had to be extracted from the corpus. The problem of sense indications also arose, especially when consulting dictionaries edited in England or the United States. A given item could exist in Canada with a meaning different from that proposed in a dictionary describing American or British varieties of English. A case in point is *Big O*, a compound used in Canadian English (especially in and around Montreal) as a nick-name for the Olympic stadium (and not to mean "orgasm" as the *Random House Dictionary of the English Language*, 1987 and the *Random House Webster's College Dictionary*, 1991

would have it). Some items could also be specific to a given variety of English, like *big bickies*, an Australian compound equivalent to *big bucks* (*Cambridge International Dictionary of English*, 1995). Dictionary sources thus have the potential to provide us with culturally biased information that might contradict actual meaning and usage in Canadian varieties of English. Finally, word spellings are often variable (e.g., *biggy*/*biggie*). We therefore had to check the corpus in order to choose which spelling of a given word is more representative of Canadian English.

2.3. Corpora

As discussed above, the number of occurrences for *big* in the corpus was huge. A first search for big in *QUEENS*, *WSJ* and *ECP* yielded 1 435, 30 665, and 78 046 matches, respectively.[2] Two different types of searches were thus performed: one search for specific collocations, compounds, derived forms and fixed expressions found in the dictionaries, and another frequency search based solely on the corpus (as described earlier). The first helped decide which items found in the dictionary would be retained. As an example, a search for the expression *to be too big for one's boots/breeches/britches/Xs* resulted in the following (total) matches, presented in Figure 1. This search confirmed Canadian usage of the expression and its spelling, and thus justified the inclusion of *too big for one's britches* under *big*.

Figure 1

Corpus matches for *to be too big for one's boots/breeches/britches*.

QUEENS: no matches

WSJ:
202347244,Committee -- too big for his overalls, some say -- is us..
74803425,s that may be too big for their britches. </s> <s> There ..

ECP:
424111684,hink she got too big for her britches." ID NUMBER: ..
483402659, getting way too big for her britches," said Los Angeles..
327652438, that he got too big for his motorcycle boots. He was ..
105311433, Has Gotten Too Big For Its Jockstrap, would lend some ..
26807924, se they got too big for their britches. It may not be ..
486704063in "Don't get too big for your britches." So the idea th..

The frequency search was also very revealing, because it provided us with a number of items that, even though frequent in the corpus, were not present in the dictionary sources. Examples of these are *big and small* (collocation), *big-city* in its adjectival use (compound), and *big shoes to fill* (fixed expression).[3] Collocations and compounds from the corpus which were not found in the dictionary sources but were nevertheless retained for the entry are marked with an asterisk in Appendix II. Although frequency searches in TEXTUM do not usually reveal fixed expressions,[4] results can hint at them. For example, the search resulted in 334 matches for *big thing* in *ECP*, a further refinement on this search revealed that a number of these matches contained the expression *the next big thing*. The search is illustrated in Figure 2 (repeat sentences have been omitted in the printout).

FIGURE 2

Corpus matches (reduced to 10) for *the next big thing* in the ECP corpus.

"big thing ": 334 matches
next fby "big thing ": 56 matches (print 30)

291486028,he country's "next big thing," 19-year-old Surrey native..
551411008,e of becoming next BIG THING Alison Mayes SOUTHAMSTAR NE..
571260652, a recovering Next Big Thing, an erstwhile rock star who..
674174424,He's been the next big thing and that old guy. He's been..
21157263,touted as the Next Big Thing and the putative kings of C..
501559190,oking for the next big thing beyond the personal compute..
745901479,really be the next big thing? BILL BROWNSTEIN GAZETTE S..
88092783,ing for the "next big thing." But even the festival's..
441095195,s Hollywood's next big thing. But he still comes across ..
396009120,uting as the "next big thing." But I left it to this w..
[…]

Thus we had to refine even further the frequency search by verifying the context around *big thing*, and not taking at face value the fact that *big thing* was a common collocation (or compound) in the ECP corpus, but rather expanding the context around discovering that it was in fact a subpart of a larger expression.

As can be seen, on the one hand, the corpus search can help us decide which items to retain from the dictionary sources. In addition, it can also be used to decide whether to include in the *BCD* any additional information not found in the dictionary sources. The final result of this cross-indexing is illustrated in Appendix III, where a number of collocations, fixed expressions, compounds and derived forms are identified in various sources, be they dictionary, corpus or both. An additional advantage of the corpus frequency search is that it provides us with information about representative structures found with the headword, even if they are not idiomatic. For example *big enough* may be a frequent collocation in the corpus. We might decide to include a free combination in the entry illustrating this fact as the *BCD* provides typical examples of language in use. Finally, the corpus has the particular advantage of highlighting typically Canadian forms and structures, thus providing us with information about the frequency of use of certain expressions, compounds and collocations in Canadian usage, in addition to information about specific meaning or spelling which these items might have in Canadian varieties of English.

3. Discussion

When creating an entry for a headword in the *BCD*, our first step is to consult what information other dictionaries (Canadian or otherwise) have included under the same headword. Dictionaries provide us with a wealth of information with regards to the syntactic and morphological structures in which we will find the headword, as well as information about compounds and collocations related to it. In addition, when a large number of dictionaries agree on a particular structure, meaning or even spelling for a given word, we can presuppose that this information is valid, although this is not always the case, as has been seen above. One possible drawback of a "dictionary-only" search for headword related information is that the written sources

may be static and not adapt as quickly to changing language use as other sources. Dictionaries might even preserve meanings and structures that are obsolete, extremely specialized, or arcane. Another potential flaw of this type of approach is that the dictionary usually reflects the language variety of the editors or the public it has been created for. Thus, British and American editions might miss or ignore Canadian usage information that would be useful to our task.

Figure 3

Corpus matches for *big wall* in WSJ

"big wall ": 39 matches
184513447,s beyond that big wall. </s>   <s> Now an ..
212088370,ment of three big Wall Street arbitragers. </s> <..
164437187,all the other big Wall Street bond desks' prices ..
134530108,l orders from big Wall Street brokers and institu..
174963090,in which many big Wall Street concerns, including..
229824084,ied against a big Wall Street firm. </s> ..
90526271,trager at one big Wall Street firm. </s> <s> "If ..
164431873, and give one big Wall Street firm a stunning adv..
177019467,sterminded by big Wall Street firms. </s> <p..
97612816, interest the big Wall Street firms, and they mus..
[…]
119270129,Monday -- for big Wall Street firms to raise a $1..
211527315,ng by several big Wall Street firms will soothe f..
90484155,Co. and other big Wall Street houses took heavy h..
16960922,sy involves a big Wall Street investment bank and..
43900751, <s> The last big Wall Street movie was "Rollover..
134519907,rn Caution As Big Wall Street Outfits Dominate St..
220735804, coattails of big Wall Street players, particula..
124956441,s trader at a big Wall Street securities firm. </..
119257248,utives at one big Wall Street securities firm wer..
62360560,ecent losses, big Wall Street securities firms sa..
67193106,ring were the big Wall Street securities firms, w..
147536233,> <s> Not all big Wall Street underwriters agree ..

Conversely, corpora provide us with a quantity of material regarding syntactic structures, collocations, compounds, derived forms and different meanings related to the headword. On the one hand, it helps us decide what information should be preserved from the written sources. On the other, it provides us with data about usage that is missing from these same sources, thus enabling us to add information in order to make a dictionary that is more representative of the Canadian varieties of English. In particular, the frequency searching is a useful tool when dealing with a word that is highly frequent in the corpus. This type of search enables us to sift quickly through a large amount of data and to pinpoint potentially important, relevant and interesting structures. However, a data base frequency search cannot be done blindly without subsequent verification of the results. For example, the frequency search in WSJ gives us 39 matches for the combination *big wall*. A subsequent verification of this result reveals that *big wall* is not a compound or a collocation. In fact it is usually found to be part of a larger noun phrase with the structure *big Wall Street X* (usually

big Wall Street firms). Results from this expanded search are illustrated in Figure 3 above.

A final disadvantage of the frequency-based search is that, in the case of frequent words like *big*, it can produce too much data without much context. The lexicographer must then go back to the written sources in order to decide which results are the most pertinent and what is superfluous. Finally, the database is also biased, in that it represents one type of language medium (written) and a limited number of styles (journalistic, literary) and interests (commerce, politics, science, etc.). The lexicographer must therefore make judgments (also biased) on what information is appropriate to include in the entry. The lexicographer must be acutely aware of personal and corpus biases, in order to tailor the entry to the expected public.

4. Conclusion

The use of databases and written sources for the creation of lexical entries is complementary. Traditional written sources provide background information on different sense divisions and common structures and compounds in which we should expect to find the headword. Databases confirm, improve on, or contradict this information and allow the lexicographer to make judgements on what to include in the dictionary entry. The data-based frequency search is especially useful when the headword is of high frequency in the corpus. It allows for a quick glance at possible syntactic and morphological structures that are common for a given headword. It attests or invalidates usage and structures proposed by the dictionary sources. This type of search is very fast and can save a lot of time in the search for the different structures of interest in the corpus. However, since the frequency search is machine-based, it is "dumb," i.e., it does not distinguish compounds, collocations, derived forms and fixed expressions from any other type of structure that might occur frequently by chance. A verification of the frequency search must therefore be carried out before this information is integrated into the dictionary entry. Even though the frequency search must be verified post-hoc, it can be very fruitful in the search for samples that are representative of local or national language varieties. Thus the corpus search, to the extent that it is available, should be used as a tool by any lexicographer intent upon devising a dictionary that is representative of language in use.

ACKNOWLEDGEMENT

This work was funded by SSHRC major collaborative research initiative funding awarded to Roberts and collaborators (1999).

NOTES

1. Unfortunately, PAT did not allow for searches of most frequent words or phrases *preceding* the headword.
2. These numbers include inflected, derived and compound forms written as one word or with a hyphen.
3. No derived forms were found that were not listed in the dictionary sources. However, some of the forms found in the dictionaries (*biggety/biggity* and *bigly*) were not found in the corpus.
4. However, this occurred with *big*: *big shoes to fill* was found in this manner.

REFERENCES

Atkins, B.T. Sue and Rundell, Michael (2007): *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.

Clas, André (1994): Collocations et langues de spécialité. *Meta*. 39(4):576-580.

Cowie, Anthony P. (1994): Phraseology. *In*: Ronald E. Asher and Seumas Simpson, eds. *The Encyclopedia of Language and Linguistics*. Vol 6. Oxford: Pergamon, 3168-3171.

Fawcett, Heather (1988): *PAT 3.3: Users Guide*. Oxford: Oxford University Press.

Fillmore, Charles(1991): "Corpus linguistics" or "Computer-aided armchair linguistics." *In*: Jan Svartvik, ed. *Directions in Corpus Linguistics, Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991*. Trends in Linguistics: Studies and Monographs 65. Berlin: Mouton de Gruyter, 35-66.

Hanks, Patrick (2008): The lexicographical legacy of John Sinclair. *International Journal of Lexicography.* 21(3):219-229.

Moon, Rosalind (2008): Sinclair, phraseology and lexicography. *International Journal of Lexicography.* 21(3):243-254.

Roberts, Roda P., Auger, Pierre, Boisvert, Lionel, Bossé-Andrieu Jacqueline, Clas, André, Cormier, Monique C. and Cummins, Sarah (1998): *Projet de lexicographie comparée du français et de l'anglais au Canada (Dictionnaire canadien bilingue)*. SSHRC major collaborative research initiative.

Roberts, Roda P. (1998): Lexicographie bilingue du français et de l'anglais au Canada. OLF. Visited on May 5th, 2009, <http://brancusi.cc.uottawa.ca/articles-en.htm>.

Roberts, Roda P. and Clas, André (2001): Should Translators Trust their Bilingual Dictionaries? Paper presented at the *XV^e congrès mondial de la FIT* (Mons, August 6-10, 1999).

Sinclair, John (1985): Lexicographic Evidence. *In*: Robert Ilson, ed. *Dictionaries, Lexicography, and Language Learning.* Oxford: Pergamon, 81-94.

Sinclair, John (1991): *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Smadja, Frank (1993): Retrieving collocations from text. *Computational Linguistics.* 19(1):143-177.

Vinay, Jean-Paul, Daviault, Pierre and Alexander, Henry (1962): *Dictionnaire canadien, français-anglais, anglais-français* (édition abrégée). Toronto: McClelland and Stewart.

APPENDIXES

1. English dictionaries used as sources for *big*, in alphabetical order by source code.

ACTIV	*Language Activator* (1994) Harlow: Longman.
BBI	*BBI Combinatory Dictionary of English* (1986) Morton Benson, Evelyn Benson and Robert Ilson, John Benjamins: Philadelphia.
CAMBR	*Cambridge International Dictionary of English* (1995/2000, electronic version) Cambridge: Cambridge University Press.
COCO2	*Collins Cobuild English Dictionary* (1995) London: Harper Collins.
COD	*Concise Oxford Dictionary* (1990) 8th ed. Oxford: Oxford University Press.
COLL	*Collins English Dictionary* (1986) Patrick Hanks, ed. Glasgow: Collins.
FUN	*Funk and Wagnalls Canadian College Dictionary* (1989) Toronto: Fitzhenry and Whiteside.
GAGE	*Gage Canadian Dictionary* (1983) Toronto: Gage.
LONG	*Longman Dictionary of Contemporary English* (1987) Harlow: Longman.
NEL	*ITP Nelson Canadian Dictionary of the English Language* (1997) Toronto: ITP Nelson.
OALD	*Oxford Advanced Learner's Dictionary* (1989) 4th ed. Oxford: Oxford University Press.
OXCAN	*Canadian Oxford Dictionary* (1998) Toronto: Oxford University Press Canada.
OXR	*Oxford Reference Dictionary* (1986) Joyce M. Hawkins, ed. Oxford: Oxford University Press.
PEN	*The Penguin Canadian Dictionary* (1990) Thomas M. Paikeday, ed. Markham/Mississauga: Penguin Books Canada/Copp Clark Pitman.

RH *The Random House Dictionary of the English Language* (1987) 2nd ed. Stuart Berg Flexner and Leonore Crary, eds. New York: Random House.

RHWEB *Random House Webster's College Dictionary* (1991) Robert B. Costello, ed. New York: Random House.

2. A sampling of *big* derived forms, collocations, compounds, and fixed expressions taken from the ECP.

Forms marked with an asterisk were included in the entry for *big* (adjective) while not found in the dictionary sources. Forms marked with a pound sign were found in dictionary sources.

ECP

>> signif.-500 «big»
12: 513 matches, text=big-time #
16: 445 matches, text=big-league #
36: 312 matches, text=big-name #
59: 204 matches, text=big-screen *
60: 203 matches, text=big-ticket #
69: 176 matches, text=big-city *
72: 173 matches, text=big-budget *
100: 127 matches, text=bigger and better
135: 102 matches, text=big-money #
155: 91 matches, text=big-band *
199: 77 matches, text=big-ticket items #
207: 74 matches, text=big-game *
219: 69 matches, text=biggest and most
261: 58 matches, text=bigger and bigger
264: 58 matches, text=biggies #

>> signif.-400 «big «1: 1077 matches, text=big deal #
2: 685 matches, text=big business #
3: 655 matches, text=big enough
5: 560 matches, text=big three #
6: 559 matches, text=big part
7: 536 matches, text=big difference
9: 453 matches, text=big leagues #
10: 440 matches, text=big bucks #
11: 428 matches, text=big hit
14: 422 matches, text=big time #
15: 410 matches, text=big city #
16: 408 matches, text=big game #
19: 374 matches, text=big brother #
20: 366 matches, text=big band #
21: 350 matches, text=big money #
22: 340 matches, text=big screen #
24: 334 matches, text=big thing *
25: 329 matches, text=big picture #
26: 326 matches, text=big in (to be) *
27: 312 matches, text=big question
29: 306 matches, text=big apple #
31: 297 matches, text=big o #
33: 259 matches, text=big winne
35: 251 matches, text=big step
37: 238 matches, text=big bang #
38: 229 matches, text=big man #
41: 216 matches, text=big win
43: 212 matches, text=big fan
44: 209 matches, text=big brothers #
45: 203 matches, text=big guy #
47: 196 matches, text=big mistake
48: 196 matches, text=big issue
49: 195 matches, text=big news #
50: 192 matches, text=big blue
51: 187 matches, text=big trouble
52: 175 matches, text=big factor
53: 175 matches, text=big ben #
56: 163 matches, text=big guns #
59: 160 matches, text=big ones #
60: 160 matches, text=big chunk
67: 146 matches, text=big break *

3. Examples of collocations, fixed expressions, compounds and derived forms identified in the dictionary entries and corpus for *big*.

Corpus data are provided in the three rightmost columns, where a "✓" indicates over ten matches.

	Dictionaries																Corpus		
	ACTIV	BBI	CAMBR	COCO2	COD	COLL	FUN	GAGE	LONG	NEL	OALD	OXCAN	OXR	PEN	RH	RHWEB	QUEENS	WSJ	ECP
Collocations																			
big boy		✓	✓		✓							✓		✓			3	3	✓
big news														✓			2	✓	✓
big sister		✓	✓	✓	✓				✓		✓	✓		✓	✓	✓	7	3	✓
the bigger the better			✓														Ø	2	✓
Fixed Expressions																			
big deal!			✓	✓	✓	✓			✓		✓			✓	✓		1	4	✓
to be big with child					✓	✓		✓	✓			✓		✓	✓		Ø	Ø	Ø
to be too big for one's boots/breeches/ britches			✓		✓	✓			✓		✓	✓		✓			Ø	1	5
what's the big deal?									✓								Ø	7	✓
what's the big idea?			✓											✓	✓		Ø	Ø	Ø
Compounds																			
Big Apple			✓	✓	✓	✓						✓		✓	✓	✓	8	✓	✓
big-bang theory			✓	✓	✓	✓	✓		✓	✓	✓				✓	✓	Ø	1	Ø
big boys			✓														Ø	✓	✓
Big Brother		✓	✓		✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	7	✓	✓
big business		✓	✓	✓	✓	✓		✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓
big cat				✓					✓			✓					Ø	3	✓
big city				✓					✓								✓	✓	✓
Big Daddy/big daddy					✓	✓									✓	✓	Ø	5	✓
big deal			✓	✓	✓	✓			✓	✓	✓	✓		✓	✓	✓	✓	✓	✓
Big Dipper/big dipper			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓			✓	✓	2	1	✓
bigeye										✓					✓	✓	Ø	Ø	Ø
big-hearted			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	1	Ø	✓
bigheartedly										✓					✓	✓	Ø	Ø	Ø
bighorn					✓	✓	✓	✓		✓			✓	✓	✓	✓	Ø	✓	✓
big leagues			✓					✓		✓		✓		✓	✓	✓	1	✓	✓
big mama															✓		Ø	6	4
big money			✓	✓	✓	✓			✓	✓	✓	✓					8	✓	✓
big shot	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	1	✓	✓
Big Sister/big sister		✓	✓									✓			✓		Ø	✓	✓
big smoke/Big Smoke			✓		✓	✓						✓					Ø	Ø	✓
big spender			✓						✓			✓					Ø	✓	✓
Big Three					✓	✓					✓	✓					8	✓	✓
big-ticket adj				✓						✓		✓		✓	✓	✓	Ø	✓	✓
big-time adj				✓		✓			✓	✓		✓		✓	✓	✓	2	✓	✓
big time adv				✓								✓					1	✓	✓
Derived words																			
biggie/biggy			✓	✓		✓		✓	✓			✓			✓	✓	2	✓	✓
biggish	✓			✓	✓	✓	✓					✓			✓	✓	Ø	Ø	7
biggity/biggety							✓								✓	✓	Ø	Ø	Ø
bigly						✓	✓			✓							Ø	Ø	Ø
bigness					✓	✓	✓	✓	✓		✓	✓		✓	✓		3	✓	✓
bigs												✓			✓	✓	Ø	✓	✓