Is this problem giving you trouble? A corpus-based examination of the differences between the nouns *problem* and *trouble*

Somjit Jirananthiporn
Department of English, Faculty of Arts,
Chulalongkorn University
E-mail: jsomjit55@yahoo.com

Abstract

This study investigates two near-synonymous nouns, *problem* and *trouble*, with a focus on differences in their frequencies, distribution patterns across text types, and collocations with verbs and adjectives. The data was extracted from the Corpus of Contemporary American English (COCA) and analyzed both quantitatively and qualitatively. The overall frequency ranking found *problem* to occur more often in all the text types, with spoken and academic discourse topping the list and fiction at the bottom. In contrast, *trouble* is most commonly used in fiction and least used in academic prose. The disparity in their occurrences in different registers was somewhat in line with findings about the verb collocates of the synonym pair. Further analysis found *trouble* to collocate more often with verbs that are characteristic of conversations pervasive in fiction, i.e. phrasal verbs, modals, contractions. Observation about adjective collocates revealed that because *problem* has a less specific meaning than *trouble*, it occurs more freely with various adjectives. Finally, in terms of pedagogical implications, the results suggest that the two words be taught with emphasis on different lexico-grammatical aspects.

Keywords: *trouble, problem*, near-synonyms, text types, collocations

1. Introduction

Vocabulary teaching is a significant part of foreign language instruction. Vocabulary plays a fundamental role in the development of language skills. This is especially true of English, which "has an impressively large lexicon" (Minkova & Stockwell, 2006, p. 461). Although the total size of the language's vocabulary is immeasurable, unabridged dictionaries of English typically contain an estimated number of entries ranging between 300,000-450,000 (Minkova & Stockwell, 2006, p. 462). It is a great consolation that to function adequately in English, consensus has it that learners need to acquire little more than 3000 words, selected by language experts based on their frequencies and usefulness (The Oxford 3000). Still, even among these fundamental words, a number of EFL learners might already find themselves struggling when encountering synonyms, which English is widely known to be rich in (Crystal, 2003). If what Edmonds & Hirst (2002) claim is true that even native speakers of a language can find it difficult to use synonyms "with invariable precision," (p.108), then

how daunting the task of mastering synonyms must be for EFL learners, like those in a Thai classroom environment. Based on the experience of this present study's researcher, the mere task of helping Thai learners to acquire new words is already painstaking and slow, let alone to distinguish between synonyms.

However, it has been demonstrated over the past decades that a corpus, a collection of "machine-readable texts, which is deemed an appropriate basis on which to study a specific set of research questions" (McEnery and Hardie, 2012: p. 1), can serve as a valuable resource for linguists and English language instructors to systematically explain uses of synonymous words in communication (see Section 2.3 for discussion of sample studies). With a focus on English synonyms that are included in the general 3000-word list, this present study is aimed at investigating two near-synonymous nouns, *problem* and *trouble,* the use of which Thai learners are prone to confuse, as illustrated in the following.

a. *I have no sleeping *trouble*.
b. I have no sleeping *problem*.
c. *Are you in a deep *problem* with your roommate?
d. Are you in deep *trouble* with your roommate?

Such errors involving collocations and grammatical patterns as shown in "a" and "c" might stem partly from the fact that, according to a bilingual dictionary, the two nouns have the same equivalent in the Thai language. However, errors of this kind continue to surface, particularly in students' writing, even with a monolingual dictionary. *Merriam Webster Learner's Dictionary* online, for example, by defining the two nouns in terms of each other, implies that they can be used interchangeably.

problem: something that is difficult to deal with; something that is a source of trouble, worry, etc.

trouble: a problem, worry, difficulty, etc. or a situation causing this

(Online *Merriam Webster Learner's Dictionary*)

The advent of online language corpora has proved a great boon to EFL classroom teaching. Instead of using a paper dictionary of collocations, researchers have at their fingertips this digital resource when they want to examine collocational behaviors of so-called near- synonyms. To investigate these near-synonyms, both of which are ranked among the general 3,000 word list available from the online *Oxford Advanced Learner's Dictionary* (OALD),

this study has been conducted using a corpus-based approach to investigate the following research questions:

1. How do *problem* and *trouble* differ in terms of their frequencies and patterns of distribution across different text types?;
2. How do *problem* and *trouble* differ in their collocations with verbs and adjectives?

Based on a corpus linguistic perspective, which highlights a combination between quantitative and qualitative aspects in the description of language use, the first research question addresses the quantitative aspect of word usage whereas the second one is qualitative. Details about English synonymy and corpus linguistic studies of near-synonyms are provided in Section 2. This is followed by descriptions of the data and descriptive tools employed to answer the research questions in Section 3. Then, results from an analysis of corpus data are reported and discussed in Section 4 before concluding remarks and relevance to the pedagogical context are given.

2. Literature Review

In this section theoretical concepts about synonymy and corpus linguistic descriptions of lexical semantics and uses are spelled out. Then sample corpus-based studies of English near- synonyms are discussed in relations to the scope of the present study.

2.1 Synonymy

Synonymy is a fundamental concept discussed in most studies investigating semantic relations of words (Harvey, 2006, cited in Phoocharoensil, 2010). It is of Greek origin and is made up of two parts *syn* meaning "alike" and *onym* meaning "name" (Murphy, 2010, p. 110). Thus, according to Palmer (1997, p. 59-63) the term is defined as "sameness of meaning."

Synonyms are usually divided into two main types: absolute synonyms and near-synonyms. The former are words characterized by "complete interchangeability", thereby substitutable in all possible contexts (Kearns, 2006, p. 558). They are generally agreed to be practically non-existent. Rare examples of this strict type of synonymy are technical terms, such as those of plants, animals and chemicals and dialectal variants such as *groundhog* (AmE) and *woodchuck* (BrE). The rarity is largely attributable to language change, which either forced an absolute synonym to be displaced or to embrace a new nuance of meaning (Clark 1992, cited in Edmonds, P. and Hirst, G. (2002), p. 107).

Near-synonyms, on the other hand, are quite easy to find. They refer to words that can be used interchangeably in some but not all contexts. Some examples are the words *fake* and *false*, which overlap in the sense of "not genuine", and therefore are substitutable in "He was arrested for carrying a ***false/fake*** ID." However, when *false* literally means "incorrect" as in "I have no idea if the statement is true or ***false***", replacing *false* with *fake* will clearly change the meaning and violate collocational constraints. As Phoocharoensil (2010) put it, these examples assert that "synonyms can be interchangeably used where their meanings overlap, but where a meaning is beyond the shared area, one cannot substitute for the other" (p. 227).

Near-synonyms not only vary in terms of denotation and collocation. Phoochaeroensil, in the same article, also described the variations of near-synonyms in other aspects: stylistics or degree of formality (e.g. peril VS danger), connotations (e.g. smile VS sneer) or grammatical patterns as illustrated in the examples below:

a. The team of psychiatrists found that he is ***mad/insane***. (mentally ill)
b. Lisa is very ***mad*** at Tim now. (angry)
c. *Lisa is very ***insane*** at Tim now.

To explain different uses of such near-synonyms as problem and trouble systematically with empirical evidence, a number of concepts and tools developed in corpus linguistics will be employed in the present study. These are explained below.

2.2 Corpus linguistics and English near-synonyms

Because a corpus contains a large number of naturally-occurring texts, it is possible to investigate uses of the English language in terms of patterns and variations. One of the most important characteristics of corpus linguistics is a combination of quantitative and qualitative perspectives on language descriptions (Biber et al, 1998). The quantitative side of corpus analysis features frequency data. With regards to near-synonyms, the degree of formality and stylistics mentioned in 2.1 above as distinguishing features of near-synonyms can be linked to the notions of frequency and distribution patterns of a lexical item across different text types. As well illustrated by various works, most notably by the *Longman Grammar of Spoken and Written English* (Biber et al, 1999), uses of lexical and grammatical features in English are marked by text registers, including conversation, fiction, newspaper and academic prose. For instance, it has been shown that conversations are characterized through a number of grammatical features such as phrasal verbs, modals and verb contractions, which are rarely found in other text types. On the other hand, academic prose tends to be full of derived words formed by

affixation. While this shows that lexical and grammatical items characterize text types, it has been observed that repeated occurrences of lexical and grammatical features in a particular register have become part of their usage profile (Stubbs, 2015). This concept has been applied to a number of studies on synonyms, e.g. Kjellmer (2003), Gries and David (2007) and Arnold (2012). On the basis of this thesis and empirical evidence, an examination of frequencies and distribution patterns of *problem* and *trouble* will be conducted to explore their differences in quantitative terms.

On the qualitative side, central to an analysis of corpus data are patterns of co- occurrence among lexical items. Findings from various corpus studies bring out one of the significant phenomena in language: "co-selection", i.e. words habitually occurring in repeated forms in a verbal environment (cf. Sinclair 1991, 2004). This has an important implication for lexicography and the study of English lexis. That is, the choice of a word is not an entirely independent selection but depends in part upon other words that the author has (not) chosen to use. Word meanings are therefore a result of the ways in which a lexical item co-occurs with other words, a notion termed by Sinclair as "units of meaning". In light of this, subtle differences between near-synonyms can be uncovered through an examination of their co- occurrence patterns. To operationalize this theoretical concept, four types of lexical co- occurrence have been pointed out by Sinclair (2004) to constitute the usage profile of a lexical item:

1. *Collocation* is the co-occurrence of words. It can also refer to the relationship a lexical item has with items that appear with greater than random probability in its context (Hoey, 1991);
2. *Colligation* is the co-occurrence of words at the grammatical level. It involves co- occurrence between a search word and a particular grammatical category. Sinclair (2004) illustrated this concept with the phrase *naked eye*, which collocates with *with* and *by*. It can therefore be described as having a preposition as its colligation;
3. Semantic preference is the co-occurrence of words at the semantic level. It refers to the co-occurrence between a search word and lexical items that share a semantic feature. For example, the phrase *naked eye*, which is found to co-occur with such words as *see*, *detect*, *study*, *spotted*, *evident* and *obvious*, can be described as having semantic preference for visibility ;
4. Semantic prosody is co-occurrence at the pragmatic level. According to Sinclair (1996 p.87), semantic prosodies are "evaluative or attitudinal and are used to express the speaker's approval (good prosody) or disapproval (bad prosody) of

> whatever topic is momentarily the object of discourse." In the case of *naked eye*, the phrase tends to co-occur with words showing difficulty in seeing, including *small*, *faint* and *weak*. The phrase is thus considered having the semantic prosody of expressing difficulty.

These different forms of co-occurrence reflect how lexical items, though similar in meaning, can have different patterns of co-occurrence at some level, which in turn points to differences in their usage profile. This concept will be adopted in an analysis of *problem* and *trouble* in the present study.

2.3 Previous corpus-based studies on synonymy

Many corpus-based studies on synonyms have been conducted, some of the most recent of which are briefly presented in this part. First, Phoocharoensil (2010) studied five verbs that are synonyms in English, i.e. *ask, beg, plead, request,* and *appeal*. In addition to examining data from three learner dictionaries: *The Oxford Advanced Learner's Dictionary, The Longman Dictionary of Contemporary English,* and *The Cambridge Advanced Learner's Dictionary,* the researcher also analyzed concordance lines from a corpus of Time in 1995, with a focus on the meanings, collocations and grammatical patterns of the five verbs studied. Their tendency for being dialect-specific as well as their stylistic differences regarding formality were also explored. The findings from comparing the data from the two main sources revealed that the corpus-based data provided additional possible grammatical patterns not illustrated in the dictionary. However, no dialectal differences were found and data showed that *request* and *appeal* are more likely found in a formal context than the other three verbs.

Using a similar approach to Phoocharoensil (2010), Thamratana (2013) also conducted a corpus- based study of verb synonyms. The research examined five synonyms: *reduce, decrease, diminish, dwindle* and *decline* in terms of meanings, grammatical patterns and formality degree. The data was drawn from *The Oxford Advanced Learner's Dictionary, The Longman Dictionary of Contemporary English,* and COCA. The study found that although these five near-synonym verbs share identical core meaning, some different grammatical patterns and collocates were detected. The adjectives decrease and decline were observed to be relatively more formal while *decrease* was found to be more commonly used in academic discourse than *reduce, decline, diminish* and *dwindle*.

Like Phoocharoensil (2010), Hoffman (2014) studied six adjective synonyms in English -- *nice, kind, lovely, friendly, gorgeous* and *pleasant* --, with the purpose of gaining new insights into these words that might ultimately help improve their definitions in dictionaries. First of all, Hoffmann compared

the similarities and differences in the definitions of these adjectives from three dictionaries: *The Macmillan Dictionary, The Collins American Dictionary and The Merriam Webster Dictionary*. The researcher then analyzed their noun collocations extracted from COCA and grouped them into lexical patterns. Also, using the corpus, Hoffmann investigated the six adjectives' patterns of distribution across the five texts and observed their stylistic variations. Comparing the information gleaned from the analysis of the data of the two sources, the researcher found that some of the definitions of the six synonymous adjectives in the dictionary are inadequate and an improvement based her findings was implied.

Cai (2012), in his Master's thesis, examined seven near-synonymous adjectives: *awesome, excellent, fabulous, fantastic, great, terrific,* and *wonderful*. Drawing on data from COCA, the researcher examined and compared the noun and adverb collocates with these adjectives to find out about their differences in meanings, uses and connotations. For example, great, due to its more general and weaker sense, was found to be more commonly used than the other adjectives, while *fabulous, fantastic* and *terrific* are more positive. In addition, the overall frequencies of the adjectives and their collocates also revealed about their distribution patterns across COCA's five text types.

On a smaller research scale, Aroonmanakun (2015)'s study supplies additional proof that corpora can be a powerful tool in learning about English synonyms. The investigation was confined to two adjectives fast and quick and how they differ with regard to their nominal collocations. The nouns analyzed were limited to the top 100 collocates which immediately follow the adjectives. The findings revealed that *fast* and *quick* do not always collocate with the same nouns. For example, ***drink*** is used more commonly with quick than fast whereas ***attack*** collocates better with *fast* than *quick*. In cases where they share the same collocates, the meanings are not the same. That is, whereas *fast* indicates the property of an action, quick signifies the manner of an action, e.g. a quick ***learner*** VS a fast ***learner***.

Chung's study (2011) is aimed at investigating the similarities and differences between two synonyms: *create* and *produce* in terms or verb forms and meanings. The data was drawn from two corpora -- The Brown Corpus and the Frown Corpus and were compared to the British National Corpus. The results found that the two verbs occur most commonly with the bare infinitive and *–ed* forms while other similarities and differences could be observed from analyzing the semantic features of the products denoted by the objects that follow the verbs. For example, the verbs were found to have two overlapping meanings: 'bring into existence/cause to happen, occur, or exist' and 'create or manufacture a man-made product', while the objects following *create* and *produce* differ. Create tends to be used with objects denoting less quantity and a

higher possibility of creativity, while *produce* is *usually* used with objects denoting greater quantity and a low level of creativity.

Despite minor differences in their methodology, all the previous studies on near-synonyms above unanimously prove how using corpora can help enhance EFL learning, especially in the area of lexicology. It is observed that one of the most popular corpora used by these researchers is Mark Davies' Corpus of Contemporary American English (COCA), which has also been chosen by the researcher of this present study. However, while all the previous studies were aimed at exploring both similarities and differences between/among the chosen near-synonyms, the present study emphasizes differences. Not only the scope but also the major corpus tools applied differentiate this study from the previous ones discussed above. This is explained in Section 3 below.

3. Methodology

This section describes the data and research tools used to extract it.

3.1 Data: The Corpus of Contemporary American English (COCA)

This study was conducted using data extracted from the Corpus of Contemporary American English (COCA). Although its database covers only the American variety, the corpus was chosen because of its free and adequate access, its large size of approximately 560 million tokens (collected between 1990-2017 according to corpus.byu.edu), its incorporation of both spoken and written languages and, last but not least, its relatively user-friendly search interface.

COCA was created by Mark Davies and launched on the Internet in 2008. The texts collected in the corpus date from the year 1990. Its size has grown from 400 million to approximately 560 million words today.

Davies (2010) explained that COCA is better than several other existing corpora e.g. the Oxford English Corpus (OEC), and British National Corpus (BNC) in that while the OEC and BNC are 'static' corpora, COCA is a 'monitor' corpus. By this, he means that new texts will continue to be stored in the corpus. Approximately 20 million words are added each year. The dynamic nature of the corpus is aimed at allowing users "to search the continually expanding corpus to see how the language is changing" (p.447).

The corpus was designed from the beginning to have its texts evenly divided into five genres: spoken, fiction, popular magazines, newspapers and academic journals. Even the proportions of sub-genres remain the same. This genre balance has given COCA an edge over other corpora.

As Davies put it:

> [this] means that we can compare data across different years and time periods and be quite certain that the corpus models accurately what is happening in the real world. As we have discussed at some length, this is quite different from the

> BoE or the OEC, which vary widely in genre composition from one period to the next, and which therefore provide some unreliable data in terms of frequency comparisons across time periods (2010, p. 462).

Accordingly, COCA is of particular relevance to the first research question of the present study in that it contains a variety of text types, with proportionate components, and therefore corresponds to the the first research question's interest in frequencies and distribution patterns across text types of the near-synonymous nouns.

In addition to elaborating on the corpus's strengths over corpora like the BNC or OEC, Davies also sheds light on the nature of the spoken texts on COCA, saying that they consist mainly of transcripts of unplanned speeches from both TV and radio programs. It is perhaps fair to assume that the spoken texts on COCA bear some resemblance to conversations. This is the stance that will help support certain findings in this study. The researcher, however, is aware that spoken language cannot be clearly divided from written language (Miller, 2006, p. 671), and that linguists do not equate it with the language of conversation.

3.2 Research tools:

Two major search options used to extract quantitative data and collocations from COCA are CHART and COMPARE. The CHART function provided the statistics for the nouns' patterns of distribution across the five text types: spoken, fiction, magazine, newspapers and academic. The COMPARE function was opted for, rather than another possible collocation extraction function on COCA – the COLLOCATES function. This is because COLLOCATES showed the frequency-based collocations of an individual lexical item. On the basis of the pilot experiment focusing on the top 20 collocates of each noun, the comparison between the collocation list of each noun showed that the two nouns shared several verb and adjective collocates, e.g. different word forms of the lemmas HAVE, BE and such adjectives as *real*, *big* and *serious*. This reflects their synonymous status. However, as the present study aims to investigate differences between two close synonyms, such similarities between major collocates of *problem* and *trouble* do not appear very insightful. On the other hand, COMPARE, through a statistical measure of significance and differences, enabled the researcher to "tease out" slight differences between the target words, which might not have otherwise surfaced with the use of common options. This will be demonstrated in the Results and Discussion section below.

COMPARE generated data for the comparison of the two near-synonymous nouns with regard to their verb and adjective collocates. The study focused only on collocates before the nouns. And, in order to generate data that is most relevant to the research questions of this paper, the query syntax was carefully determined. In the examination of verbs, the collocation query was set

to only two span positions to the left of the node. This was to account for the fact that *problem* is a countable noun, thereby requiring a determiner between the noun and the verb. The syntax was also appropriate because this study was limited to the singular form of the noun *problem*. For adjective collocates, the search syntax was set to one span position to the left of the node.

In the first place, the two near-synonyms' overall frequencies across different discourses were generated via the CHART function. This was in order to examine overall usage patterns of the nouns on the basis of quantitative data. The results were then compared and analyzed qualitatively. The next step involved the comparison and examination of statistically significant collocates of *problem* and *trouble*. Verb collocates were dealt with first followed by the adjectives. The same verb collocates with different inflections were counted as a lemma represented by the base form. The results extracted from CHART and COMPARE were explored to determine if there were any connections between them. The tallied lists of verb and adjective collocates were analyzed qualitatively to determine whether they differ in terms of usage and grammatical patterns as well as text types. Concordance lines and online dictionaries were consulted to confirm or reject the observations made from the analysis.

4. Results and Discussions

This section is structured according to the research questions spelled out in the introductory part, starting from the results of the quantitative analysis, followed by the investigation of collocations of *problem* and *trouble*.

4.1 Overall frequency and distribution patterns of problem and trouble

The quantitative investigation of the words *trouble* and *problem* via COCA in terms of overall frequency patterns reveals that the number of occurrences of problem is over three times greater than that of *trouble* (165,528 tokens as opposed to 49,140). This indicates that *problem* is more commonly used in communication than *trouble*. (see Table1).

Table 1: *Overall frequency and distribution patterns of* ***problem*** *and* ***trouble*** *across text types.*

	Problem		**Trouble**	
Text types	**Frequency**	**Per million**	**Frequency**	**Per million**
Spoken	52568	450.27	12737	109.10
Fiction	16652	148.88	14778	132.13
magazine	31078	246.82	9748	83.06
newspapers	27182	270.56	8850	78.32
academic	38048	341.51	3027	27.17
Total	**165528**		**49140**	

By sheer number of overall occurrences, it should then follow that problem outnumbers its near-synonym counterpart across the five text types. A query of COCA to examine the nouns' distribution patterns across the text types confirms the hypothesis. (see Table 2).

Detailed investigation gave further insight into the two nouns' frequency rank across text types.

Table 2: *Overall frequency and distribution patterns of* ***problem*** *and* ***trouble*** *across text types from largest to smallest*

	Problem			**Trouble**	
Text types	**Frequency**	**Per million**	**Text types**	**Frequency**	**Per million**
spoken	52568	450.27	**fiction**	14778	132.13
academic	38048	341.51	**spoken**	12737	109.10
magazine	31078	246.82	**magazine**	9748	83.06
newspapers	27182	270.56	**newspapers**	8850	78.32
fiction	16652	148.88	**academic**	3027	27.17
Total	**165528**		**Total**	**49140**	

As can be seen from Table 2 above, the word *problem* is found to occur most frequently in spoken discourse, and second most frequently in academic texts. The lowest frequency of the word is in fiction. An interesting point arising from this distribution pattern is that the noun *problem* occurs most frequently in two somewhat contradictory text types, spoken discourse and academic prose (cf. Conrad and Biber, 2004). This can be taken to reflect the profile of the word *problem* as a general noun in English, a noun that occurs frequently in different types of texts (Mahlberg, 2005).

On the other hand, as shown in Table 2, the distribution patterns of the noun *trouble* point in an opposite direction. It occurs most frequently in fiction, where *problem* was found to occur the least, and least frequently in academic discourse, the runner-up text type in the case of *problem*. Such a disparity suggests that although both nouns are close in their semantic properties, they are used in different contexts of communication. The word *problem* can be used generally, whether in a more formal context such as in academic prose or more informally as in conversation, while uses of *trouble* seem to be oriented towards more informal discourse. This is also manifested through the fact that the frequency rank reversal of the two nouns' register distribution pattern is greatest in fiction discourse, which appears at the top of the text type list for *trouble* and at the bottom for *problem*. Given that fiction is an expressive text, this clear-cut discrepancy seems to add further support to the observation that *problem* is more formal than *trouble*.

What is also interesting is that spoken language discourse tops the frequency ranking for *problem* and holds the second place for *trouble*. The fact that they are often used in conversations underpins this researcher's claim that

the two words can cause difficulty for Thai EFL learners as both nouns are common in the same text type. This is in addition to the problems students may encounter when consulting dictionaries, which as noted in the Introduction, define the two words using the concept of difficulty and even list *problem* as a synonym of *trouble*. Nevertheless, as will be seen in Section 4.2, a close look at their collocational patterns will shed more light on their differences.

In the next section, the stylistic issue of formality and informality raised by frequency information in this section will be further explicated in lexico-grammatical terms when we query COCA to examine how their verbal and adjectival collocates might vary according to the context used.

4.2 Analysis of verb and adjective collocates

Through the function COMPARE, verb and adjective collocates of the two near-synonyms are extracted and will be discussed in turn below.

4.2.1 Verb collocates

COCA yielded the top 100 statistically significant verb collocates for problem and 75 for trouble (See Appendix 1) The verb collocates of these two nouns display differences in their grammatical and semantic co-occurrence patterns.

4.2.1.1 Verb collocates of *problem* and *trouble*: A grammatical perspective

Upon a comparison of the lists of collocations, it is found that no words in the two lists overlap. All 100 collocates of *problem* on the list are lexical verbs, which in fact consist of 53 different verb lemmas. In contrast, of the 75 collocates of trouble, 68 tokens are lexical verbs, which were reduced to 26 lemmas, and the remaining 17 tokens are modal and primary auxiliaries in contracted and full forms, including *would, may, might, can, could, should, will, 'll, would, 'd, do, does, did, 'm*. In Sinclair's (1996) term, these auxiliaries can be seen as colligates of the noun trouble, while they do not occur significantly with *problem*. It should be noted, however, that since auxiliaries are tagged as verbs, the query syntax, which was set at two span positions before the node, allowed the mass noun *trouble* to be preceded by a lexical verb plus an auxiliary.

A closer look at the results of the tallying above provides further evidence to support the findings concerning the distribution patterns of the two nouns discussed in 4.1. On the one hand, all 100 verb collocates of *problem* are almost exclusively derived words formed with affixation (see table 3), with the exception of two phrasal verbs: *engage* (in) and *turn* (into, over, back, around). Many of these verb collocates, such as *alleviate, combat, compound, exacerbate,* and *rectify* are not found among the list of 3000 general words in English, suggesting that they are more specific to a particular

text type or domain rather than being commonly used in any register. A further investigation reveals that these single-word collocates are frequently used in academic texts, which in turn accounts for why the noun is very commonly used in the academic realm and not in fiction writing. This is not true of *trouble*'s verb collocates, however. Nine of the 26 lexical verb collocates (34.62 per cent) associated with trouble are either two-word verbs: *get* (into), *look* (for), *run*, (into), *come* (into), *head* (for), *stir* (up) and *ask* (for), or verbs with a preposition: *lead* (to) and *sound* (like). As phrasal verbs or "two-word" verbs are considered to be characteristic of conversational discourse (Biber et al., 1999, p. 408), these results are in line with the fact that fiction and spoken text types are the top two contexts where *trouble* is found in COCA.

The association of *trouble* with informal communicative style is also enhanced by its colligation with auxiliary verbs, especially modals, pointed out above. As auxiliary verbs are particularly a signal of spoken texts (Biber, 1999, p. 486), their frequent co-occurrence with *trouble*, especially those in contracted forms, helps underscore the result of the distribution pattern across the text in 4.1 that *trouble*, unlike *problem*, is used primarily and extensively in informal language. A random check of the concordance lines of *'ll* found one instance in which the node *trouble* appeared as a verb instead of a noun. However, a more careful look at the concordance lines reassures that this was an isolated case.

4.2.2.2 Verb collocates of *problem* and *trouble*: A semantic perspective

With *problem*, the researcher first sorted out 53 verb lemmas, checking the concordance lines and consulting the online dictionary to refine the categorizations. Six lemmas are removed from the list: *DIAGNOSE*, *TREAT*, *WORK*, *ENGAGE (IN)*, *USE*, and *LIE* for the following reasons. The first two are used when problem means "illness" and the third WORK is used exclusively with "solving a math problem". *ENGAGE* and *USE* are followed by a noun phrase with *problem* as a modifier e.g. "problem solving", "problem types". Lastly, *LIE* is an intransitive verb used in the subject–verb inversion structure. The removal of these verbs results in the remaining 47 lemmas, 45 of which can be grouped according to their semantic properties, illustrating the semantic preference pattern of *problem*. The members of each group are demonstrated in Table 3 below.

1. The group "finding a solution to a problem" is the largest group of verb collocates of *problem*, consisting of 19 verbs with the meaning of "finding a way to deal with a problem/difficult situation (in order to solve the situation).
2. "The neutral", the second largest group contains 18 words, which are lumped together because they can be said to carry a neutral

effect. That is, they do not render the overall meaning positive or negative. Most of the verbs convey activities aimed at clarifying or learning more about a problem.

3. The "initial action" group contains four verbs that convey the concept of "to start to deal with a difficult situation or to think about a problem or a situation and decide how you are going to deal with it (with no commitment of finding a solution)".
4. The "causing difficulty" group consists of two verbs which share the meaning of "putting someone in a difficult situation."
5. The "make worse" group consists of two verbs meaning "to worsen the problem/situation"

With these categories, the verbs 'remain' and 'become' do not fall into any semantic group, as they do not express related meanings but are related in terms of their grammatical categories as copular verbs.

Table 3: ***Problem****'s verb collocates categorized according to semantic preference*

1	2	3	4	5
alleviate, attack, combat, control, correct, cure, eliminate, fight, fix, handle, manage, overcome, rectify, reduce, repair, remedy, resolve, solve, tackle,	*acknowledge, analyze, assess, consider, define, discover, discuss, describe, explain, identify, ignore, illustrate, investigate, recognize, represent, study, turn, understand*	*address, answer, approach, confront*	*present, pose*	*compound, exacerbate*

In this part, we investigate trouble's 26 verb lemmas. (See Appendix 1 for the full list) Since trouble can also be a verb, it is necessary to browse the concordance lines for some of the suspicious verbs in the list to ensure that they are followed by trouble as a noun. The result is four collocates-- begin, *come, continue, seem* -- had to be removed from further analysis. *Borrow* was also excluded because "borrow trouble" is an idiomatic expression whose meaning does not retain the original meaning of *trouble*. *Love* was also taken off the list. All its 24 occurrences in the concordance lines are part of the title of the movie "I Love Trouble". Lastly, the verb *take* was also removed because it was found to be part of the idiom "take the trouble to do something", in which case the noun 'trouble' conveys the meaning of 'effort' rather than 'difficulty.'

Upon examination of the remaining 19 verb lemmas (see Table 4), it was found that their semantic property does not play as big a role as in the analysis of *problem*. Most of the verbs convey more or less the meaning "get oneself or someone else into a difficult situation" or "give oneself/someone else problems," e.g. This suggests that while *problem* tends to have various semantic preferences and be used in various discourses, ranging from "solving the difficulty," to "dealing with difficulty" and "causing difficulty", the noun *trouble* tends to be mainly used in one context, i.e. "difficulty occurs to someone". However, the verb collocates of *trouble* seem to be distinguishable in terms of grammatical patterns associated with them, whereas *problem's* verb collocates are all transitive verbs with one similar pattern: : "verb + determiner + *problem*". The 19 verbs are grouped according to grammatical patterns in the table below.

Table 4: *Grammatical patterns of **trouble**'s verb collocates*

(1) Phrasal verbs	(2) Intransitive verb + preposition	(3) Transitive Verbs	(4) Others forms
ask (for) *head* (for) *get* (in/into) *look* (for) +*trouble* *run* (into) *stir* (up)	*lead* (to) + *trouble* *sound (like)*	*expect* *invite* *like* *mean* +*trouble* *sense* *smell* *spell* *want*	*start* +having/causing/making +*trouble* *have* +determiner (no, some) +*trouble* +(V ing) give +object pron. (us/him/her) +*trouble*

This has pedagogical implications for teaching the phraseology of the two nouns: for *trouble*, attention should be paid to grammatical patterns of the noun. On the other hand, when teaching the noun *problem*, more weight should be given to its collocates and various semantic preferences.

4.2.2 Adjective collocates:

To compare the adjective collocation results of the two noun near-synonyms, the query syntax was set at one position to the left of the node. Like with noun collocate findings, the table comparing the two lists of collocates showed that adjective collocates of *problem* outnumber those of *trouble*, as can be seen in Appendix 2. That is, the COCA search came up with a list of 100 statistically significant adjective collocates for *problem* and only seven for *trouble.*

In the case of *problem*, of all its 100 collocates, five adjectives were eliminated from the analysis because the concordance details revealed them to qualify not the noun *problem* itself but the entire noun phrase "problem solving" (*creative, individual, collaborative,* and *cooperative*), and "problem scores" (*total*). Moreover, one duplication of the collocates was found in the list:

"long-standing" and "longstanding". They appeared separately but the concordance check revealed that they are simply a matter of different spellings. Hence, they were treated as one type.

The remaining 94 words were then examined and categorized into four groups according to their semantic relationship. The first and largest group consists of 41 adjectives. They either signify the degree of severity/importance or the extent of the problem. Examples of these adjectives are *fundamental, difficult, growing, complex, central, important,* and *larger,* all of them found among the top ten on the list. Some adjectives, such as *unique, inherent, chicken- and-egg,* may not fit in easily with this category but they do convey the sense of seriousness of a problem. This is illustrated by the following:

- Perhaps these is no area within schools and between schools where a cooperative effort is more important than the curriculum. This presents a **unique problem** to school districts.

- although ecologists are very good at the generation of hypotheses, it can be very difficult to decide which hypotheses are worthy of strong consideration.
  This **inherent problem** can be characterized by a statement that has become known as Phaedrus's law:

- I want to get people excited about biking so that it becomes part of our social currency. " For that to happen, public transportation has to solve its big **chicken-and- egg problem**. Most people don't want to use trains, buses or bikes unless they're really convenient, but most cities aren't willing to spend enough to make these services convenient until enough people start using them.

The second largest group includes 33 adjectives that are related to the subject area of the problem (e.g. *social, mathematical, homeless, cultural, ethical,* and *behavioral*), the geographical area and the people the issue concerns (e.g. *national, regional, worldwide, Palestinian, American, Kurdish and Jewish*). The third group conveys the meaning of the generality/specificity of a difficult situation (*single, unique, classic, individual,* and *universal*). The fourth group contains nine adjectives that provide temporal details of a problem: *recurring, original, age-old, on-going, perennial, short-term, persistent, long- standing* and *final*.

The remaining seven adjectives are lumped together in the miscellaneous group because they do not lend themselves to any of the first three semantic categories and they themselves do not appear to share related semantic properties: different, opposite, related, interesting, and given.

Table 5: 94 collocates of *problem* categorized by semantic property

Severity	*fundamental, difficult, growing, complex, central, important, larger, underlying, pressing, systemic, critical, tough, urgent, broader, primary, unresolved, pervasive, biggest, vexing, large, basic, common, intractable, simple, crucial, chief, key, essential, slight, widespread, insurmountable, complicated, challenging, inherent, easy, unsolved, principal, nagging, thorny, significant small*
Subject	*national, social, Palestinian, American, mathematical, homeless, technical, cultural, Kurdish, ethical, general, behavioral, societal, Jewish structural, nuclear, philosophical, medical, regional, nationwide worldwide, scientific, mind-body, chicken-and-egg, theoretical, clinical historical, black, Indian, moral, strategic, theological, three-body*
Generality/ specificity	*single, unique classic, individual, universal*
Temporal details	*recurring, original, age-old, on-going, perennial, short-term, persistent. long-standing, final*
Miscellaneous	*different, opposite, related, interesting, given*

With regards to *trouble*, the word *back* in the list was eliminated since a concordance examination found the word to be a noun, not an adjective. In addition, the adjective *foul*, which, despite its highest ratio score, was eliminated because its exclusive occurrence in a sports context seems to limit its usefulness for EFL learners. Therefore, only five adjectives are left for further analysis: *deep, desperate, extra, double, financial*.

Like those of *problem*, these five collocates can be seen as semantically related in some ways. The adjective collocates *deep* and *desperate*, *extra* and *double* convey the degree of severity, and *financial* specifies the subject matter of the trouble.

Based on the above categorization, the adjectives commonly used with *problem* and *trouble* share two semantic preferences: "severity" and "subject". This suggests their status as close synonyms. At the same time, the noun *problem*'s semantic preference for "temporal details" is not shared by *trouble*. It can therefore be stated that, compared with their verb collocates, the

two close synonyms *problem* and *trouble* are only slightly different in their adjective collocates.

Besides, the discrepancy between the length of the two adjective collocate lists gives us further insight how the two near-synonym nouns differ. Unlike *problem*, it is much less common for *trouble* to be immediately preceded by an adjective. An examination of the context of occurrences of the two nouns reinforced the observation that *trouble* is usually preceded by a preposition (e.g. *of*, in), an adverb (*into*) or a determiner (*no*, *any*, *the*) and very rarely by an adjective, while for *problem*, it is quite frequent to have an adjective immediately preceding it. This is illustrated through the sample concordance lines below (figures 1 and 2).

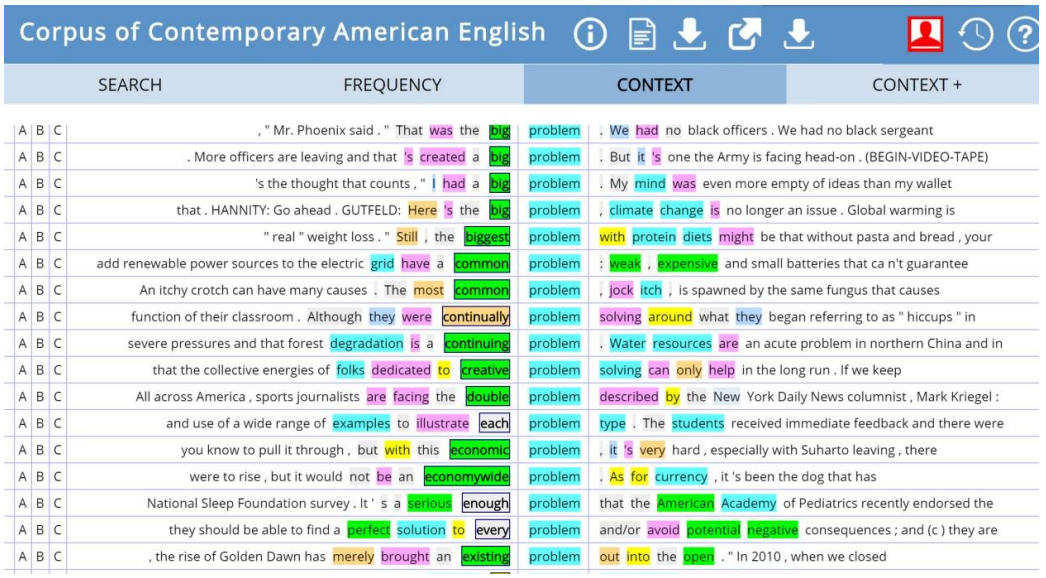

Figure 1: Concordance lines for *problem*

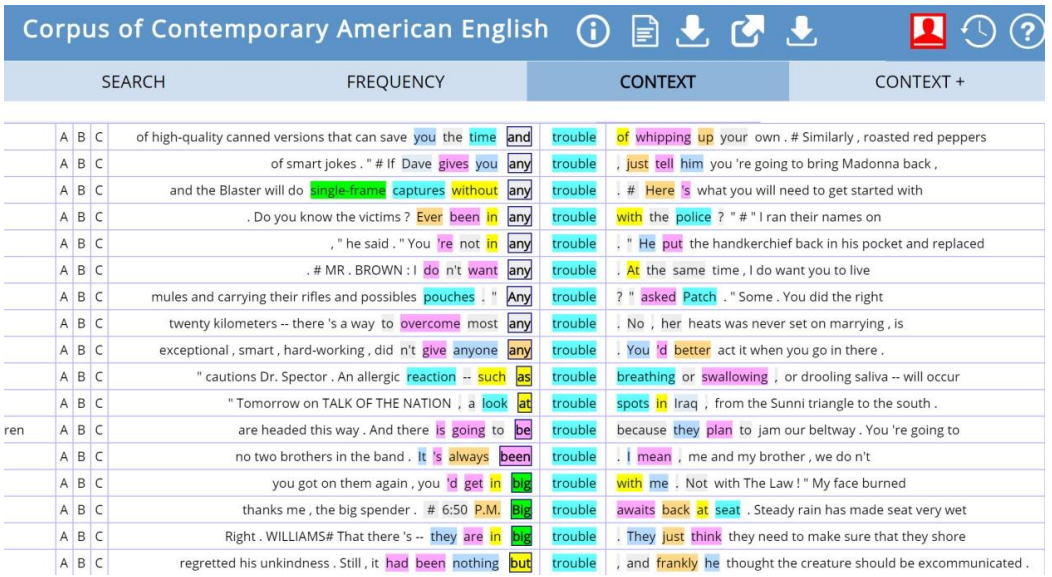

Figure 2: Concordance lines for *trouble*

To account for what causes such a difference between problem and trouble, it is worth looking into the semantic features of the two near-synonyms. It may be useful to take a look again at the most common definition of each of the nouns below.

problem: something that is difficult to deal with; something that is a source of trouble, worry, etc.

trouble: a problem, worry, difficulty, etc. or a situation causing this

(Online Merriam Webster Learner's Dictionary)

On the one hand, the indefinite pronoun "something," in the definition of problem carries a sense of "lack of specificity". A sense of difficulty is thus not inherently entailed in "something" but rather in the action of coping with it. As a result, *problem* is open to a qualifying word specifying what the "difficult something" is about and the degree of difficulty/or importance of that "something" involved in finding a solution for. For *trouble*, a sense of difficulty is integral and the need to find a solution for it is not incorporated in the word. Hence, the noun is usually used abstractly to mean difficulty itself.

5. Conclusion

This study is aimed at uncovering, by means of a corpus-based approach, the differences between the two near-synonymous nouns *problem* and *trouble* with regard to their frequencies, distribution patterns, verb and adjective collocations, colligations and semantic preferences. The data was obtained from COCA and analyzed both quantitatively and qualitatively. The analysis yielded a number of insights that benefit EFL instruction.

First of all, through distribution patterns across text types, *problem* is revealed to be a much more widely used word than *trouble,* with the former's frequencies far outnumbering the latter across the five genres. The frequency information also revealed *trouble* to be favored in spoken discourse rather than in formal written genres, represented by academic texts. This observation is further supported by the analysis of the target nouns' verb collocates. That is, *problem* was found to be associated with verbs that are characteristic of written discourse whereas *trouble* tends to occur with verbs used in spoken language. *Problem*'s statistically significant verb collocates were larger in number, consisting almost exclusively of multi- syllable lexical verbs while those of *trouble* comprise mostly phrasal verbs and one-syllable words as well as quite a few primary and modal auxiliaries in both complete and contracted forms. Another marked difference between the two nouns seems to be of relevant benefit to EFL instruction. The result concerning the nouns' semantic preferences and grammatical patterns suggests that when teaching *problem* emphasis should be placed on the first, while with *trouble* emphasis should be placed on the latter.

With regard to adjective collocates, the much greater number of *problem*'s statistically significant collocates positioned right before the noun compared to *trouble*'s suggests a subtle implication for the EFL context. The finding suggests that *problem* used without any adjective can be vague, due to the word's inherent neutral meaning. This is not true of *trouble,* which explains why it is much less commonly preceded by an adjective. Awareness of the semantic properties of the two words is therefore recommended when dealing with advanced learners.

It should be noted that the study's findings were largely influenced by the use of the COMPARE function to extract data. Consequently, the top lists of collocates generated were obviously different compared to those produced by the mere COLLOCATE option. For example, the resulting short list of statistically significant adjective collocates for *trouble* might not be in line with samples provided by a dictionary. Besides, it is worth remarking that since the study of the noun *problem* is limited to the singular form only, its findings concerning collocation behaviors should not be presumed to be true of its plural form.

Despite these limitations, the study did give insights that are valuable to EFL instruction and learning and, to certain extent, beyond what a dictionary tells. A dictionary may be based on corpora from a certain period of time. But given that language is dynamic and corpora continue to grow to reflect the endless changes, a study such as this has its intrinsic benefits.

Works Cited

Arnold, C. (2012). *Synonymy and Register. A Corpus-based Study in American English.*Bachelor Thesis. Johannes Gutenberg University Mainz.

Aroonmanakun, V. (2015). Quick or Fast: A Corpus-based study of English synonyms. *Language Education and Acquisition Research Network* (LEARN) Journal. *8*(1), 53-62.

Biber, D., Johansson, S. Leech, G., Conrad. S., & Finegan E. (1999). *Longman grammar of spoken and written English*. London: Longman.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge Univ. Press.

Cai, J. (2012). *Is it "great" enough? A corpus-based study of "great" and its near synonyms*. Master Thesis. Ball State University.

Chung, Siaw-Fong. (2011). A corpus-based analysis of "create" and "produce". *Chang Gung Journal of Humanities and Social Sciences, 4*(2) (October 2011), 399-425.

Clark, E. V. 1992. Conventionality and contrast: Pragmatic principles with lexical consequences. In Lehrer, A., and Kittay, E. F. (Eds.), *Frames, Fields, and Contrasts: New Essays in Semantic and Lexical Organization*. (pp. 171–188). Lawrence Erlbaum.

Conrad, S. M., & Biber, D. (2004). "The frequency and use of lexical bundles in conversation and academic prose" *Lexicographica - International Annual for Lexicography / Internationales Jahrbuch für Lexikographie* 20: 56-71

Corpus of Contemporary American English. Retrieved 2018. Retrieved March 29, 2018, from https://corpus.byu.edu/coca/.

Crystal, D. (2003). *The Cambridge encyclopedia of the English language*. Cambridge: Cambridge University Press.

Davies, M. (2010). The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing, 25*(4). 447-464.

Edmonds, P. & Hirst, G. (2002). Near-synonymy and lexical choice. *Computational Linguistics 28*(2), 105-44.

Gries, S. Th., & David, C.V. (2007). This is kind of / sort of interesting: variation in hedging in English. *Towards Multimedia in Corpus Studies 2* (Studies in Variety, Contacts and Change in English). ed by Päivi Pahta, Irma Taavitsainen, Terttu Nevalainen & Jukka Tyrkkö Helsinki. VARIENG. http://www.helsinki.fi/varieng/series/volumes/02/gries_david/

Hoey, M. (1991) *Patterns of Lexis in Text*. Oxford: Oxford University Press

Hoffmann K. (2014). A corpus-based analysis of the near-synonyms 'Nice', 'Kind', 'Lovely', 'Friendly', 'Gorgeous' and 'Pleasant. 122228 BA Paper – Usage-based approach to linguistic investigations. Accessed

March 28, 2015. https://www.academia.edu/7541436/A_corpusbased_analysis_of_the_nearsynonyms_nice_ki nd_lovely_friendly_gorgeous_a nd_ple asant_

Kearns, K. (2006). Lexical Semantics. In Aarts B., & McMahon, A. (Eds.), *The Handbook of English Linguistics*. (pp.692-723). Oxford: Blackwell Publishing.

Mahlberg, M. (2005). *English General Nouns: A Corpus Theoretical Approach*. Amsterdam, John Benjamins.

Miller, J. (2006). Spoken and Written English. In Aarts B., & McMahon, A. (Eds.), *The Handbook of English Linguistics*. (pp.670-691). Oxford: Blackwell Publishing.

Minkova, D. & Stockwell, R. (2006). English words. In Aarts B., & McMahon, A. (Eds.), *The Handbook of English Linguistics*. (pp. 461-482). Oxford. Blackwell Publishing.

Murphy, M. L. (2010). *Lexical Meaning*. Cambridge University Press, Cambridge.

Phoocharoensil, S. (2010). A Corpus-based study of English synonyms. International Journal of Arts and Sciences, 3(10), 227-245.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Sinclair, J., & Carter, R. (2004). *Trust the Text Language, Corpus and Discourse*. Florence: Taylor and Francis.

Stubbs, M. (2002). *Words and phrases: Corpus studies of lexical semantics*. Malden: Blackwell.

Stubbs, M. (2015). The textual functions of lexis. In Groom, N., Charles, M., & John, S. (Eds.), *Corpora, Grammar and Discourse*. (pp. 97-116). Amsterdam; Philadelphia: Benjamins.

Thamratana, P. (2013). *A corpus-based analysis of reduce, decrease, diminish, dwindle and decline*. Independent study. Language Institute, Thammasat University, Bangkok, Thailand.

Appendix 1: Full lists of verb collocations

	WORD 1 (W1): PROBLEM (3.37)				
	WORD	W1	W2	W1/W2	SCORE
1	SOLVE	2972	2	1,486.0	441.1
2	SOLVED	537	0	1,074.0	318.8
3	SOLVING	525	0	1,050.0	311.7
4	RESOLVE	263	0	526	156.2
5	ADDRESSING	204	0	408	121.1
6	SOLVES	177	0	354	105.1
7	CORRECT	162	0	324	96.2
8	ADDRESS	778	3	259.3	77
9	FIX	641	3	213.7	63.4
10	APPROACH	101	0	202	60
11	ATTACK	98	0	196	58.2
12	ADDRESSED	92	0	184	54.6
13	FIXING	91	0	182	54
14	PRESENTS	90	0	180	53.4
15	POSED	86	0	172	51.1
16	DISCUSS	80	0	160	47.5
17	TACKLE	156	1	155	46
18	ALLEVIATE	71	0	142	42.2
19	FIXED	69	0	138	41
20	OVERCOME	136	1	136	40.4
21	DEFINE	67	0	134	39.8
22	ADDRESSES	66	0	132	39.2
23	COMBAT	66	0	132	39.2
24	IDENTIFIED	64	0	128	38
25	PRESENTED	64	0	128	38
26	BECOMES	127	1	127	37.7
27	TACKLING	57	0	114	33.8
28	EXACERBATED	51	0	102	30.3
29	REMAINS	97	1	97	28.8
30	DEFINING	48	0	96	28.5
31	DISCUSSED	48	0	96	28.5
32	RESOLVING	48	0	96	28.5
33	PRESENT	92	1	92	27.3
34	REPAIR	45	0	90	26.7
35	ELIMINATE	88	1	88	26.1
36	COMPOUNDS	44	0	88	26.1
37	EXACERBATES	44	0	88	26.1
38	TREAT	44	0	88	26.1
39	EXACERBATE	86	1	86	25.2
40	ATTACKING	41	0	82	24.3
41	APPROACHED	40	0	80	23.7
42	DESCRIBE	38	0	76	22.6
43	RECOGNIZE	75	1	75	22.3
44	STUDY	75	1	75	22.3
45	APPROACHING	37	0	74	22
46	RESOLVED	37	0	74	22
47	ENGAGE	36	0	72	21.4
48	EXACERBATING	36	0	72	21.4
49	REDUCING	36	0	72	21.4
50	REDUCE	70	1	70	20.8
51	CORRECTING	34	0	68	20.2
52	ILLUSTRATES	34	0	68	20.2
53	REMEDY	67	1	67	19.9
54	ACKNOWLEDGED	33	0	66	19.6
55	EXPLAINED	33	0	66	19.6
56	COMPOUND	63	1	63	18.7
57	MANAGE	31	0	62	18.4
58	CONTROL	29	0	58	17.2
59	RECTIFY	29	0	58	17.2
60	EXPLAIN	56	1	56	16.6
61	ANALYZE	28	0	56	16.6
62	ATTACKED	27	0	54	16
63	ILLUSTRATE	27	0	54	16
64	STUDIED	27	0	54	16
65	STUDYING	27	0	54	16
66	CONSIDER	53	1	53	15.7
67	CURE	53	1	53	15.7
68	IGNORE	53	1	53	15.7
69	DEFINED	26	0	52	15.4
70	INVESTIGATE	26	0	52	15.4
71	TURN	25	0	50	14.8
72	POSES	97	2	48.5	14.4
73	COMPOUNDED	46	1	46	13.7
74	DESCRIBING	23	0	46	13.7
75	ENGAGED	23	0	46	13.7
76	REMAIN	23	0	46	13.7
77	TACKLED	23	0	46	13.7
78	USING	23	0	46	13.7
79	LIES	89	2	44.5	13.2
80	BECAME	88	2	44	13.1
81	BECOMING	44	1	44	13.1
82	ANSWER	22	0	44	13.1
83	FIGHT	22	0	44	13.1
84	HANDLED	22	0	44	13.1
85	CONSIDERED	43	1	43	12.8
86	CORRECTED	21	0	42	12.5
87	CONFRONTING	21	0	42	12.5
88	ELIMINATED	21	0	42	12.5
89	RECOGNIZING	21	0	42	12.5
90	DISCOVERED	41	1	41	12.2
91	COMPOUNDING	120	3	40	11.9
92	UNDERSTANDING	40	1	40	11.9
93	DIAGNOSING	20	0	40	11.9
94	ACKNOWLEDGING	20	0	40	11.9
95	REPRESENT	20	0	40	11.9
96	ASSESS	19	0	38	11.3
97	PRESENTING	19	0	38	11.3
98	WORKING	19	0	38	11.3
99	DISCUSSING	37	1	37	11
100	ACKNOWLEDGES	18	0	36	10.7

	WORD 2 (W2): TROUBLE (0.30)				
	WORD	W2	W1	W2/W1	SCORE
1	ASKING	152	0	304	1,024.0
2	RAN	137	0	274	923
3	SPELL	105	0	210	707.4
4	GETTING	432	3	144	485.1
5	GOTTEN	142	1	142	478.3
6	LL	71	0	142	478.3
7	STIR	67	0	134	451.4
8	SPELLS	56	0	112	377.3
9	STIRRING	53	0	106	357.1
10	RUNNING	43	0	86	289.7
11	SPELLED	36	0	72	242.5
12	HEADED	71	1	71	239.2
13	RUN	185	3	61.7	207.7
14	GETS	168	3	56	188.6
15	GIVES	26	0	52	175.2
16	LOOKING	205	4	51.3	172.6
17	LEAD	47	1	47	158.3
18	HEADING	22	0	44	148.2
19	MIGHT	156	4	39	131.4
20	LED	19	0	38	128
21	EXPECTING	37	1	37	124.6
22	D	105	3	35	117.9
23	SMELLED	17	0	34	114.5
24	LOOK	30	1	30	101.1
25	VE	112	4	28	94.3
26	SOUNDS	13	0	26	87.6
27	RUNS	25	1	25	84.2
28	BORROW	12	0	24	80.8
29	INVITE	12	0	24	80.8
30	WOULD	284	13	21.8	73.6
31	GET	1393	64	21.8	73.3
32	GIVING	62	3	20.7	69.6
33	LEADS	10	0	20	67.4
34	HEAD	10	0	20	67.4
35	M	319	16	19.9	67.2
36	EXPECTED	19	1	19	64
37	TAKEN	73	4	18.3	61.6
38	COULD	229	13	17.6	59.3
39	RE	905	55	16.5	55.4
40	MEANT	65	4	16.3	54.7
41	START	48	3	16	53.9
42	SEEM	16	1	16	53.9
43	GIVE	63	4	15.8	53.1
44	MAY	201	14	14.4	48.4
45	GAVE	39	3	13	43.8
46	LOVE	24	2	12	40.4
47	SENSING	23	2	11.5	38.7
48	SMELL	11	1	11	37.1
49	WILL	175	17	10.3	34.7
50	LOOKS	10	1	10	33.7
51	HAVING	2549	281	9.1	30.6
52	TAKES	18	2	9	30.3
53	LOOKED	15	2	7.5	25.3
54	WANT	139	20	7	23.4
55	BEGAN	27	4	6.8	22.7
56	DID	98	16	6.1	20.6
57	MEAN	83	14	5.9	20
58	SENSED	23	4	5.8	19.4
59	SEEMED	17	3	5.7	19.1
60	MEANS	45	8	5.6	18.9
61	ASK	11	2	5.5	18.5
62	TOOK	69	13	5.3	17.9
63	AM	42	8	5.3	17.7
64	SHOULD	41	8	5.1	17.3
65	LIKE	10	2	5	16.8
66	SENSES	10	2	5	16.8
67	DOES	100	21	4.8	16
68	CAME	19	4	4.8	16
69	STARTED	30	7	4.3	14.4
70	DO	74	18	4.1	13.8
71	COME	12	3	4	13.5
72	COMES	23	6	3.8	12.9
73	CONTINUE	10	3	3.3	11.2
74	WERE	448	146	3.1	10.3
75	CAN	103	34	3	10.2
76	NEED	27	10	2.7	9.1
77	EXPECT	27	10	2.7	9.1
78	CAUSE	412	166	2.7	9
79	MAKE	266	116	2.3	7.8
80	TAKE	100	51	2	6.6
81	KNEW	24	13	1.8	6.2
82	HIT	11	6	1.8	6.2
83	SIGNAL	11	6	1.8	6.2
84	WANTED	11	6	1.8	6.2
85	MAKING	110	62	1.8	6
86	ARE	817	472	1.7	5.9
87	HAD	3632	2119	1.7	5.9
88	CAUSING	203	119	1.7	5.7
89	TAKING	25	15	1.7	5.6
90	GIVEN	25	16	1.6	5.3
91	IMAGINE	12	8	1.5	5.1
92	SPOT	23	18	1.3	4.3
93	ANTICIPATE	10	8	1.3	4.2
94	BRING	31	26	1.2	4.2
95	CAUSES	47	38	1.2	4.2
96	KEEP	16	14	1.1	3.8
97	EXPERIENCED	12	11	1.1	3.7
98	MADE	36	33	1.1	3.6
99	AVOIDED	19	18	1.1	3.6
100	HAS	1011	965	1	3.5

Appendix 2: Full lists of adjective collocations

	WORD 1 (W1): PROBLEM (3.37) ADJ				
	WORD	W1	W2	W1/W2	SCORE
1	FUNDAMENTAL	452	0	904	268.4
2	DIFFICULT	289	0	578	171.6
3	GROWING	288	0	576	171
4	COMPLEX	204	0	408	121.1
5	CENTRAL	193	0	386	114.6
6	NATIONAL	172	0	344	102.1
7	SOCIAL	338	1	338	100.3
8	IMPORTANT	156	0	312	92.6
9	LARGER	265	1	265	78.7
10	PALESTINIAN	118	0	236	70.1
11	UNDERLYING	221	1	221	65.6
12	AMERICAN	102	0	204	60.6
13	PRESSING	96	0	192	57
14	SYSTEMIC	94	0	188	55.8
15	PRACTICAL	90	0	180	53.4
16	CRITICAL	89	0	178	52.8
17	SPECIFIC	155	1	155	46
18	MATHEMATICAL	77	0	154	45.7
19	DIFFERENT	151	1	151	44.8
20	HOMELESS	75	0	150	44.5
21	TOUGH	74	0	148	43.9
22	TECHNICAL	144	1	144	42.7
23	OPPOSITE	72	0	144	42.7
24	URGENT	71	0	142	42.2
25	BROADER	70	0	140	41.6
26	PRIMARY	65	0	130	38.6
27	SINGLE	65	0	130	38.6
28	RELATED	64	0	128	38
29	UNRESOLVED	60	0	120	35.6
30	CREATIVE	114	1	114	33.8
31	PERVASIVE	57	0	114	33.8
32	CULTURAL	55	0	110	32.7
33	INTERESTING	55	0	110	32.7
34	KURDISH	54	0	108	32.1
35	RECURRING	54	0	108	32.1
36	BIGGEST	1467	14	104.8	31.1
37	ORIGINAL	52	0	104	30.9
38	ETHICAL	103	1	103	30.6
39	AGE-OLD	51	0	102	30.3
40	GENERAL	98	1	98	29.1
41	ONGOING	98	1	98	29.1
42	BEHAVIORAL	49	0	98	29.1
43	VEXING	49	0	98	29.1
44	LARGE	48	0	96	28.5
45	BASIC	284	3	94.7	28.1
46	PERENNIAL	46	0	92	27.3
47	SOCIETAL	46	0	92	27.3
48	COMMON	451	5	90.2	26.8
49	JEWISH	45	0	90	26.7
50	STRUCTURAL	88	1	88	26.1
51	GIVEN	44	0	88	26.1
52	NUCLEAR	44	0	88	26.1
53	PHILOSOPHICAL	44	0	88	26.1
54	SHORT-TERM	44	0	88	26.1
55	INTRACTABLE	83	1	83	24.6
56	SIMPLE	41	0	82	24.3
57	MEDICAL	240	3	80	23.7
58	PERSISTENT	79	1	79	23.5
59	CRUCIAL	39	0	78	23.2
60	REGIONAL	39	0	78	23.2
61	TOTAL	73	1	73	21.7
62	CHIEF	36	0	72	21.4
63	KEY	141	2	70.5	20.9
64	ESSENTIAL	35	0	70	20.8
65	NATIONWIDE	35	0	70	20.8
66	SLIGHT	35	0	70	20.8
67	WIDESPREAD	69	1	69	20.5
68	LONG-STANDING	34	0	68	20.2
69	WORLDWIDE	33	0	66	19.6
70	INSURMOUNTABLE	32	0	64	19
71	SCIENTIFIC	32	0	64	19
72	COMPLICATED	62	1	62	18.4
73	CHALLENGING	31	0	62	18.4
74	INHERENT	31	0	62	18.4
75	MIND-BODY	31	0	62	18.4
76	LONGSTANDING	30	0	60	17.8
77	UNIQUE	30	0	60	17.8
78	FINAL	29	0	58	17.2
79	CHICKEN-AND-EGG	29	0	58	17.2
80	THEORETICAL	29	0	58	17.2
81	EASY	28	0	56	16.6
82	INDIVIDUAL	28	0	56	16.6
83	CLINICAL	27	0	54	16
84	HISTORICAL	27	0	54	16
85	UNSOLVED	27	0	54	16
86	BLACK	26	0	52	15.4
87	CLASSIC	26	0	52	15.4
88	PRINCIPAL	51	1	51	15.1
89	COLLABORATIVE	25	0	50	14.8
90	INDIAN	25	0	50	14.8
91	NAGGING	25	0	50	14.8
92	THORNY	25	0	50	14.8
93	UNIVERSAL	25	0	50	14.8
94	MORAL	99	2	49.5	14.7
95	SIGNIFICANT	293	6	48.8	14.5
96	STRATEGIC	24	0	48	14.2
97	THEOLOGICAL	24	0	48	14.2
98	THREE-BODY	23	0	46	13.7
99	SMALL	137	3	45.7	13.6
100	COOPERATIVE	22	0	44	13.1

	WORD	W2	W1	W2/W1	SCORE
1	FOUL	149	0	298	1,003.8
2	DEEP	438	16	27.4	92.2
3	BACK	10	0	20	67.4
4	DESPERATE	15	2	7.5	25.3
5	EXTRA	14	2	7	23.6
6	DOUBLE	82	17	4.8	16.2
7	FINANCIAL	269	70	3.8	12.9
8	CONSIDERABLE	31	11	2.8	9.5
9	IMPENDING	10	5	2	6.7
10	WORSE	27	19	1.4	4.8
11	BAD	44	32	1.4	4.6
12	LEGAL	126	95	1.3	4.5
13	LITTLE	332	319	1	3.5
14	FURTHER	42	41	1	3.5
15	RECENT	13	13	1	3.4
16	FEMALE	10	12	0.8	2.8
17	GREAT	56	121	0.5	1.6
18	TERRIBLE	53	131	0.4	1.4
19	BIG	846	2149	0.4	1.3
20	DEEPER	32	89	0.4	1.2
21	SERIOUS	573	1618	0.4	1.2
22	WORST	16	56	0.3	1
23	ECONOMIC	47	176	0.3	0.9
24	POSSIBLE	18	70	0.3	0.9
25	CONSTANT	11	48	0.2	0.8
26	POLITICAL	79	404	0.2	0.7
27	REAL	458	2372	0.2	0.7
28	POTENTIAL	65	391	0.2	0.6
29	NEW	42	254	0.2	0.6
30	MECHANICAL	10	69	0.1	0.5
31	OTHER	58	423	0.1	0.5
32	ONLY	131	1202	0.1	0.4
33	BIGGER	29	584	0	0.2
34	PARTICULAR	18	373	0	0.2
35	HUGE	22	570	0	0.1
36	MAIN	16	593	0	0.1
37	MAJOR	31	1282	0	0.1
38	BIGGEST	14	1467	0	0