A CORPUS-BASED STUDY OF THE EFFECTS OF COLLOCATION, PHRASEOLOGY, GRAMMATICAL PATTERNING, AND REGISTER ON SEMANTIC PROSODY

by

TIMOTHY PETER MAIN

A thesis submitted to

the University of Birmingham

for the degree of

DOCTOR OF PHILOSOPHY

Department of English Language
and Applied Linguistics
College of Arts and Law
The University of Birmingham
December 2017

Abstract

This thesis investigates four factors that can significantly affect the positive-negative semantic prosody of two high-frequency verbs, CAUSE and HAPPEN. It begins by exploring the problematic notion that semantic prosody is collocational. Statistical measures of collocational significance, including the appropriate span within which evaluative collocates are observed, the nature of the collocates themselves, and their relationship to the node are all examined in detail. The study continues by examining several semi-preconstructed phrases associated with HAPPEN. First, corpus data show that such phrases often activate evaluative modes other than semantic prosody; then, the potentially problematic selection of the canonical form of a phrase is discussed; finally, it is shown that in some cases collocates ostensibly evincing semantic prosody occur in profiles because of their occurrence in frequent phrases, and as such do not constitute strong evidence of semantic prosody. Finally, register-specific examination of grammatical patterning of CAUSE shows that both register and patterning can affect semantic prosody. This study shows that although positive-negative semantic prosody is an important aspect of meaning, it is potentially problematic, and any claims that a word or phrase has a positive or negative semantic prosody may require taking these factors into consideration.

Dedication

This thesis is dedicated to So-hyun and Lizzy for their endless patience and support.

“if you do not love me I shall not be loved
if I do not love you I shall not love”

Acknowledgements

I would like to thank my lead supervisor, Dr. Crayton Walker, for his patience, support, and impeccable guidance throughout the years spent preparing this thesis. Crayton's enthusiasm and dedication to bringing out the best in the work have made all the difference.

CONTENTS

List of Tables

List of Figures

Conventions and Abbreviations Used in this Thesis

SMALL CAPITAL LETTERS

- used for lemmas — all of the word forms of an item — for example, the lemma CAUSE is realized by the word forms *cause, causes, caused,* and *causing*

Italics

- used to refer to a specific word form, e.g. "Both *thing* and *things* are significant collocates of *happen.*"

- used for phrases extracted from corpus data presented in the body text of the thesis (i.e. not in concordances or tables of collocates), for example, "The concordance line contains the phrase *heat loss caused by convection* …."

Bold text (other than chapter and section headings)

- used for grammatical patterning notation. For example, the pattern **V n to-inf** is realized by *causing the spin of the earth to slow*.

- used to add emphasis to quotations, for example, "All **subsequent choices** within the lexical item relate back to prosody." Sinclair (2004c, p. 34, emphasis added)

- used for sub-corpora of the Bank of English, for example, "**newsci** comprises full issues of *New Scientist* magazine."

Numbers in parentheses following a lemma, word form, or phrase

- refer to the frequency of that item in the corpus, for example, "The short phrase *things happen* (1,497) was selected …"

Node

- refers to the word or phrase under investigation in the corpus. Collocates are referred to by their positions surrounding the node (N+1, N-1, etc.) as the following illustrates:

N-4	N-3	N-2	N-1	Node	N+1	N+2	N+3	N+4
people	with	long	noses	happen	to	be	more	intelligent

Frequency as Collocate (FaC)

- refers to the frequency of a word as a collocate of the node, either in a 4:4 List or in a specific position in the Picture, e.g. "*things* has a raw frequency of 197,244 in the BoE and an FaC of 1,497 at N-1 of *happen*."

Capitalized terminology:

- List: refers to an automatically generated list of collocates created by the Bank of English Lookup software. The List programme ranks the top fifty collocates in the 4:4 span (regardless of position) according to the statistical measure chosen by the researcher (see Chapter 3 for detailed discussion of these measures).
  i. Frequency List: collocates are ordered by raw frequency as collocate (FaC)
  ii. T-List: collocates are ordered by t-score
  iii. MI-List: collocates are ordered by MI score

- Picture: refers to an automatically generated table of collocates created by the Bank of English Lookup software. Each column of the table contains the top fifty collocates ranked by the selected statistical measure for that position relative to the node. The researcher may choose spans of 3:3 to 6:6 (see Section 5.4.2 for detailed explanation and examples).
    i. Frequency Picture: collocates are ordered by raw Frequency as Collocate (FaC)
    ii. T-Picture: collocates are ordered by t-score (Section 2.4.4)
    iii. MI-Picture: collocates are ordered by Mutual Information score (Section 2.4.4)

- Positional Frequency Table (PFT): refers to a manually generated table of collocates ranked by frequency, similar to the BoE Picture, created using Microsoft Excel (see Section 5.4.2 for detailed explanation and examples).

CHAPTER 1:
INTRODUCTION

1.1 Background and General Aim of the Thesis

This thesis is a corpus-based study of factors that affect the semantic prosodies of two frequent verbs, CAUSE and HAPPEN. This study uses corpus data from the 450-million word Bank of English (BoE) and the English Web 2013 corpus (enTenTen13) (discussed in detail in Sections 5.3.1 and 5.3.2 respectively), to show that collocation, phraseology, grammatical patterning, and register all have significant effects on the semantic prosodies of these two verbs. The general aim of the study is to show that semantic prosody is a potentially problematic phenomenon that requires further detailed study.

This chapter introduces the notion of semantic prosody and explains the rationale behind the current study. Section 1.2 discusses the origins of the term itself and outlines two major competing models of semantic prosody. Section 1.3 addresses two of the more pertinent criticisms of semantic prosody. Section 1.4 discusses the specific objectives of this thesis, and Section 1.5 outlines how the thesis is organized.

1.2 Semantic Prosody

The first use of "semantic prosody" in print is in Louw (1993, p. 158), who attributes his use of the term to personal communication with John Sinclair in 1988. Louw further refers to Sinclair's (1987, p. 155) observations of SET *in*: "The most striking feature of this phrasal verb is the nature of the subjects. In general, **they refer to unpleasant states of affairs**" (emphasis added). He notes that a small number of subjects of *SET in* are neutral, but the majority are negative, and none "is desirable or attractive." He observes: "The main vocabulary is *rot* (3),

decay, malaise, despair, ill-will, decadence, impoverishment, infection, prejudice, vicious (*circle*), *rigor mortis, numbness, bitterness, mannerism, anticlimax, anarchy, disillusion, disillusionment, slump*" (Sinclair, 1987, p. 156). Louw also cites Sinclair's (1991, p. 112) similar comments on the typical behaviour of HAPPEN: "Many uses of words and phrases show a tendency to occur in a certain semantic environment. For example, the verb *happen* **is associated with unpleasant things—accidents and the like**" (emphasis added). Louw then presents his own corpus evidence for the negative semantic prosodies of *utterly*, *days are*, *bent on*, *fine friend*(*s*), and *symptomatic of*.

Louw uses these final two items, *fine friend*(*s*) and *symptomatic of*, to illustrate a "secondary, although no less important attitudinal function of semantic prosodies" (Louw, 2000, p. 56), namely their role in the instantiation of irony in a text or in observations of a language user's possible insincerity. Essentially, where a prosodic "clash" is purposive the effect is the creation of irony, as in the corpus example he presents (Louw, 1993, p. 167), reproduced in part, below:

1. Six ninety-five at Ohrbach's. The only piece of clothing she had bought since she came to New York. "Will I shame you in front of your **fine friends**?" she said, "A dozen of my **fine friends** will come up to you tonight and ask for your telephone number, he said."

Louw (1993, p. 167) suggests that the ironic reading is confirmed both by "the repetition of the authentic example" (the first speaker uses the phrase ironically, the second does not) and by the co-selection of *shame*.

Where prosodic clash is unintentional, the effect is to reveal the language user's true feelings about the subject. Louw (1993) quotes an interview with the Director General of the British Council and argues that the Director's previously 'hidden' attitudes were unwittingly expressed through an ostensibly innocuous choice of words. Louw (1993, p. 169) quotes the director's

response to a question about how wide the network between the University of Zimbabwe and British universities is:

> Well, it's very wide. I mean, it's *symptomatic of* the University of Zimbabwe which has such a high reputation that there are fifteen links between departments in the university here and equivalent departments in all sorts of institutions, universities, polytechnics in Britain. That is a huge number of links and reflects not only the closeness with which Zimbabwean and British educators have been working but, as I say, the level of the University of Zimbabwe.

Louw (1993, p. 170) then presents a concordance of *symptomatic of* that shows "overwhelming evidence of a negative semantic prosody"; the concordance shows negative co-selections such as *something deeply wrong*, *other management inadequacies*, *deeper endemic tensions*, and so on. Following is a concordance of the first twenty-three lines (one screen of the Telnet application used to access the corpus) of *symptomatic of* (537) in the Bank of English. This short concordance appears to confirm Louw's observations.

Figure 1.1 Twenty-three lines (one screen of the BoE Telnet window) of *symptomatic of* confirming Louw's (1993) observations of negative semantic prosody

```
2.  the ultimate problem or even symptomatic of a problem in the person
3.    and white SAT scores... is symptomatic of what happens when education
4.    food chains -- all of this symptomatic of the transborder communities
5.      Unwillingness to work is symptomatic of the lack of grace
6.   of superpower relations are symptomatic of the balance of power
7.  were unfounded, but they are symptomatic of growing concern about
8. ibis's imminent extinction is symptomatic of much larger ecological
9.     on the Internet. They are symptomatic of what may become one of the
10. Shiva believes the patent is symptomatic of a far more serious form of
11.  called amyloid fibrils, are symptomatic of a number of diseases. The
12.    two hundred years ago was symptomatic of their popularity on the
13.be small things, but they are symptomatic of the lack of care taken
14.   of English in India may be symptomatic of something far more corrosive
15.         again. <p> But it is symptomatic of Gorbachev's hesitations
16.   the reservations policy is symptomatic of dangerous fragmentation of
17.  night's game which might be symptomatic of why some people find it
18.   decision to send troops is symptomatic of the conservative resurgence
19.changing his mind in public is symptomatic of what some politicians call
20.  runs out in May and that is symptomatic of a bigger problem.
21.  propaganda against Nahda is symptomatic of a new danger. Kenneth
22.     the Los Angeles riots as symptomatic of a society that's lost its
23.        but the incidents are symptomatic of the vulnerability of plants
24.     the Cubs should be in is symptomatic of all that is diseased in the
```

The observation of the overwhelming negative semantic prosody of *symptomatic of* leads Louw (1993, p. 170) to the conclusion that the Director "believes that the University of Zimbabwe badly needs assistance from Britain" despite his apparent efforts to conceal this opinion by co-selecting *high reputation*. The "prosodic clash" between the negative prosody of *symptomatic of* and the positive meaning of *high reputation* reveals to the listener/reader the Director's insincerity.

Xiao and McEnery (2006, p. 106) provide some examples of semantic prosodies; their table is reproduced in Table 1.1 below. These items — CAUSE, HAPPEN, etc. — have in common the fact that, in themselves, they do not seem to have evaluative meanings; that is, "the item does not appear to have an affective meaning until it is in the context of its typical collocates" (Xiao and McEnery, 2006, p. 107).

Table 1.1 Examples of semantic prosodies, taken from Xiao & McEnery (2006, p. 106)

Author	Negative	Positive
Sinclair (1991)	BREAK *out* HAPPEN SET *in*	
Louw (1993, 2000)	bent on build up of END up *verb*ing GET oneself *verb*ed a recipe for	BUILD up a
Stubbs (Stubbs, 1995, 1996, 2001a, 2001c)	ACCOST CAUSE FAN the flame signs of underage teenager(s)	PROVIDE career
Partington (1998)	COMMIT PEDDLE/peddler dealings	
Hunston (2002)	SIT through	

Louw (2000, p. 56) writes:

> A semantic prosody refers to a form of meaning which is established through the proximity of a consistent series of collocates, often characterizable as positive or negative, and whose primary function is the expression of the attitude of its speaker or writer towards some pragmatic situation.

Importantly, "[t]his knowledge is not necessarily either conscious or explicitly recollectable but remains part of our communicative competence[1]" (Partington, 2004, p. 132).

Although Louw credits Sinclair with both coining the term and establishing the theory of semantic prosody, it is not until the 1996 article, "The search for units of meaning" (cited here from the reprinted version in the 2004 volume *Trust the Text*), that Sinclair uses the term "semantic prosody" in print, and here he attributes the term to Louw (1993). However, by this time, Sinclair's model of semantic prosody has become somewhat more complex. In this article, he begins to lay out the structure of the "lexical item" which he describes in terms of five categories of co-selection — the item's core, collocation, colligation, semantic preference, and semantic prosody (Section 2.2 looks at these categories in detail). Sinclair (2004c, p. 34) argues:

> The semantic prosody has a leading role to play in the integration of an item with its surroundings. It expresses something close to the 'function' of the item — it shows how the rest of the item is to be interpreted functionally. Without it, the string of words just 'means' — it is not put to use in a viable communication.

Sinclair describes the five categories of co-selection in more detail in his 1998 article, "The lexical item" (again, citations here refer to the 2004 reprint in *Trust the Text*). Here, he makes an important argument, namely that the "core"— which "constitutes the evidence of the occurrence of the item as a whole" (Sinclair, 2004b, p. 141)[2] — and semantic prosody are obligatory categories of co-selection, while the others are optional. The core, he argues, is required because it is "invariable, and constitutes the evidence of the occurrence of the item as

a whole"; that is, without the core, there is nothing with which the other categories can be *co*-selected. Semantic prosody is described as obligatory because it "is the determiner of the meaning of the whole" (Sinclair, 2004b, p. 141). Sinclair (2004b, pp. 144–145) argues:

> The semantic prosody of an item is the reason why it is chosen, over and above the semantic preferences that also characterize it. It is not subject to any conventions of linguistic realization, and so is subject to enormous variation, making it difficult for a human or a computer to find it reliably. It is a subtle element of attitudinal, often pragmatic meaning and there is often no word in the language that can be used as a descriptive label for it. What is more, its role is often so clear in determining the occurrence of the item that the prosody is, paradoxically, not necessarily realized at all.

Within this broader model, Sinclair moves beyond the dichotomy of positive or negative prosody proposed by Louw (1993, 2000) and proposes a semantic prosody of "difficulty" for the lexical item *naked eye* (2004c, p. 34), and later (Sinclair, 2004b, p. 145) he argues that the item *budge* has a semantic prosody of "frustration".

The fact that semantic prosody has been defined in at least two distinct ways almost since its inception is one of the reasons that, in the abstract to her seminal article "Semantic prosody revisited", Susan Hunston (2007) characterizes semantic prosody as a "contentious term". Indeed, one of the main issues she addresses is the fact that semantic prosody has been used to refer to both "the discourse function of an extended unit of meaning, and the attitudinal meanings typically associated with a word or phrase" (Hunston, 2007, pp. 255–266). She attributes the former model to Sinclair, and the latter to Partington (2004), but, as we have seen, this positive-negative approach to semantic prosody is also put forth by Louw (1993, 2000) and has been further explored by Louw and Chateau (2010) and others (Dilts and Newman 2006; Morley and Partington 2009; Partington 1998, 2004; Stubbs 1995, 1996; Wei and Li 2014; Xiao and McEnery 2006).

Louw and Chateau (2010, p. 756), for example, are quite clear: "It is the combined association of different words having the same polarity, positive or negative (generally negative), which identifies the polarity preferentially associated with the node word or expression, and thus its semantic prosody." Hunston (2007, p. 256) argues, however, that this strictly evaluative approach "can involve taking a somewhat simplistic view of attitudinal meaning. Such meaning is often not reducible to a simple 'positive or negative'."

Like Sinclair, who initially describes the prosodies of SET *in* and HAPPEN in dichotomous terms, but then expands his model to include more specific expressions of prosodic meaning, earlier research by Stubbs (1995, 1996) tends to describe prosodies in terms of a simple positive-negative polarity: "some words (e.g. CAUSE) have a predominantly negative prosody, a few (e.g. PROVIDE) have a positive prosody" (Stubbs, 1996, p. 176). However, in later writings Stubbs' approach appears closer to Sinclair's. Stubbs (2001c, p. 65) writes: "Since they are evaluative, prosodies often express the speaker's reason for making the utterance."

Hunston is, of course, correct; evaluative meaning is not always reducible to a good/bad contrast. It could be argued, however, that in many cases the choice of how to express an item's prosody is a matter of specificity — 'negative' is simply less specific than "difficult" or "frustration" for example. Importantly, for some items, expressing the semantic prosody in more specific terms does not seem possible. The lemma CAUSE is an example of a negative semantic prosody that would be difficult to express in more specific terms. Table 1.2 shows the main evaluative collocates of CAUSE in the BoE profile arranged to show its semantic preferences, defined by Stubbs (2001c, p. 65) as "the relation […] between a lemma or word-form and a set of semantically related words, and often it is not difficult to find a semantic label for the set." Unlike *budge*, the prosody of which Sinclair (2004b, p. 144) describes as "frustrating or

irritating", CAUSE resists this level of specificity. The preferences in Table 1.2 are all indisputably negative, but no other comprehensive label seems to apply.

Table 1.2 Evaluative collocates of CAUSE in the BoE arranged to display semantic preferences

Diseases/Injuries/ Medical Symptoms	Psychological/ Emotional Distress	Social Disruptions/ Disturbances	General Damage/ Loss	Complications/ Hindrances
aids	alarm	assault	accident	delays
bacteria	anxiety	chaos	crash	difficulties
cancer	concern	conspiracy	collapse	disruption
death	confusion	controversy	damage	failure
deaths	consternation	crime	decline	lack
discomfort	distress	crisis	explosion	problem
disease	embarrassment	furore	harm	problems
diseases	offence	havoc	loss	risk
hiv	outrage	mayhem	losses	trouble
injuries	panic	storm		
injury	stress	uproar		
pain	upset			
suffering	fear			
symptoms	grief			
virus				

The five preferences shown in Table 1.2 illustrate Stubbs' (2001c, p. 66) observation:

> The distinction between semantic preference and discourse [semantic] prosody is not entirely clear-cut. It is partly a question of how open-ended the list of collocates is: it might be possible to list all words in English for quantities and sizes, but not for 'unpleasant things'. It is also partly a question of semantics versus pragmatics.

This semantic/pragmatic distinction leads Stubbs (2001c, p. 65) to briefly consider adopting the term "pragmatic prosodies", but he settles instead on "'discourse prosodies', both in order to maintain the relation to speakers and hearers, but also to emphasize their function in creating discourse coherence" (Stubbs, 2001c, p. 66).

Stubbs is not alone in expressing the need to reformulate (often subtly) and rename semantic prosody to reflect this important semantic/pragmatic distinction. Hoey (2005, p. 24) prefers "semantic association", although he is careful to note that "[t]he change of term does not

represent a difference of position between Sinclair and myself." And Hunston (2007, p. 266) writes:

> [M]y own suggestion would be that the term 'semantic prosody' is best restricted to Sinclair's use of it to refer to the discourse function of a unit of meaning, something that is resistant to precise articulation and that may well not be definable as simply 'positive' or 'negative'. I would suggest that a different term, such as 'semantic preference' or perhaps 'attitudinal preference', is used to refer to the frequent co-occurrence of a lexical item with items expressing a particular evaluative meaning.

For his part, however, Sinclair did not appear to see the necessity of bifurcating the theory into two separate and distinct models. For example, even as recently as in 2003, Sinclair continues to express the prosody of HAPPEN in terms of "good" and "bad". In "Task 14: Hidden Meanings" of *Reading Concordances*: *An Introduction*, he asks: "Can you tell whether the 'happening' is regarded as a good thing, a bad thing, or in between, neutral?" (Sinclair, 2003, p. 117). In the same exercise, he also refers to "good and bad expectations" associated with the ostensibly neutral verb *happen*, and later still he classifies all of the instances in his sample concordance of *happen* by their evaluative polarity: definitely good, probably good, neutral, probably bad, and definitely bad.

Dilts and Newman (2006, p. 233) point out: "The most common understanding [of semantic prosody] that we seem to encounter […] is that some words, or word groups, occur in contexts which are understood by the researcher to have "positive" or "negative" nuances, or prosodies." Morley and Partington (2009, p. 141), similarly contend that "corpus linguists seem to be reaching a general agreement in appreciating the good-bad, positive-negative distinction at the heart of the notion of evaluation." They contend that semantic prosody is, at its core, primarily evaluative, and that "evaluation at its most basic is a two-term system" (Morley and Partington 2009:141). They (2009:141–42) continue with a convincing argument that does not preclude a

wider focus on the pragmatic or discourse role of semantic prosodies, but which finds that the most interesting insights into language and communication are derived from semantic prosody as a positive-negative dichotomy:

> Rather than 'simplistic', we would prefer to say that the good-bad distinction is the essential simplicity at the heart of a complex system. If one loses sight of this and treats every version, every variation of goodness and badness as a separate prosody, one loses the fundamental original insight of the concept of semantic prosody, in other words, the extraordinary unifying explanatory power regarding the function of communication that evaluation and semantic prosody provide.

Bednarek (2008, p. 133) takes a similar approach, acknowledging the complexity of evaluative meaning, but recognizing that focus tends to fall on positive-negative evaluation:

> Even though most discussions of semantic prosody (apart from those by Sinclair) have predominantly focused on positive and negative attitudinal meanings, it should be kept in mind that there are many semantic prosodies that do not relate to '(un)pleasantness' and that evaluation is much more multi-faceted.

For the reasons outlined above, I have chosen to adopt the simple, though not necessarily "simplistic" notion: "words can have a specific halo or profile, which may be positive, pleasant and good, or else negative, unpleasant and bad" (Bublitz, 1996, p. 10). This approach does not preclude that other expressions of semantic prosody are possible, but it does acknowledge that the positive-negative distinction is not only usually a sufficient characterization of semantic prosody, but at times it is the only characterization possible.

1.3 Criticisms of Semantic Prosody

Perhaps the most vehement criticisms of semantic prosody have been made by Whitsitt (2005), who argues for nothing less than a complete dissolution of the entire notion. Sections 1.3.1 and 1.3.2 look at two of Whitsitt's (2005) more salient arguments and related issues. These sections show how semantic prosody researchers have responded to these criticisms and how these

potentially problematic aspects of semantic prosody are approached in the current study.

1.3.1 Criticism one: diachronic processes are not observable in synchronic corpora

Whitsitt (2005, p. 296) contends that the analogies used in Louw (1993) to suggest that an item takes on elements of meaning from a set of frequently co-selected items are faulty. His argument is founded specifically on an attack of the analogy that words acquire semantic prosodies by being "imbued" with the evaluative meanings of their most frequent collocates.

McEnery and Hardie (McEnery and Hardie, 2012, p. 139) counter this argument:

> Some of Whitsitt's criticisms are ill-founded — for instance, in our judgement his attacks on the analogies and metaphors that Louw (1993) employs in outlining the nature of semantic prosody, whether or not they are accurate in substance, do not amount to an invalidation of the concept, which must stand or fall on its own merits rather than those of the analogies used to present it.

However, it is worth looking deeper into a closely related facet of this criticism of semantic prosody, namely the apparent diachronic nature of imbued meaning. Whitsitt (2005, pp. 287–88) asks: "can the process of diachronic change be derived from the observations made of a synchronically organized corpus? The answer is no, and therefore the concept should be dismissed." Walker (2008, pp. 43–44) demonstrates that semantic prosody is likely not a diachronic process by looking to the earliest usage examples of *utterly*, UNDERGO, and CAUSE in the Oxford English Dictionary (OED). The negative prosodies associated with these three items are clearly observable in these samples, each hundreds of years old, which suggests that the prosodies are not, in fact, dependent on change in meaning over time. Walker (2008, p. 44) argues, however that "the evidence from the OED does not completely invalidate Louw's model of semantic prosody," as Whitsitt demands.

Indeed, very little, if anything, in Louw's model requires it to be diachronic. Though it is possible that Louw was incorrect about this specific aspect of semantic prosody, it is not necessary to dispense with the concept of semantic prosody entirely. Louw's (1993) only reference to the supposed diachronic nature of semantic prosody is in one short paragraph in which he discusses their potential for "instantiating irony" (see discussion above) and argues that this is only possible once a prosody has become sufficiently strong. The instantiation of irony would seem to be equally effective, however, whether the semantic prosody has built up gradually over time to a sufficient strength, or whether it was established more or less fully-formed with the item as Walker's examples in the OED suggest.

Whitsitt (2005, p. 298) is adamant, however, that any argument that meaning "flows from one group of words to another" is necessarily invalid because such flow is impossible to begin with. He (2005, pp. 296–297) argues:

> One need but consider verbs like *alleviate*, *heal*, *relieve*, *soothe*, etc., all perfect candidates for semantic prosody since they all habitually appear in the company of clearly unpleasant words, yet it seems clear that a word like *alleviate*, to take one example, certainly does not come to have an unpleasant meaning because of that company.

Morley and Partington propose a rather elegant response to this type of criticism, by briefly introducing a notation style that they argue represents a somewhat more complex system of "embedded evaluation":

> The different states of affairs can be represented notationally as [*exacerbate* [a problem]] and (*alleviate* [a problem]), where square brackets indicate 'bad', round brackets 'good', and where the outer bracketing indicates the overall evaluation and the prosody of the key item.

In both of their examples, *problem* is bad and is identified as such by the square brackets surrounding it. However, *alleviate a problem* is surrounded by round brackets, indicating that

the proposition, as a whole, evaluates positively. In practice, the majority of examples of semantic prosody would not require the added level of clarity provided by this method of notation, but it does seem to address on a theoretical level that an overarching prosody can appear to contradict collocational evidence.

Hunston counters Whitsitt's argument quite differently. She claims the middle ground between Whitsitt, who denies any possibility of intertextuality at all, and Louw, who appears to require that the semantic prosody of an item is absolute (barring only cases of irony or insincerity). Hunston (2007, p. 266) observes that the transfer of evaluative meaning to an item from its frequently co-selected items does not have to be expressed in absolute terms:

> To say that a word cannot possibly carry an attitudinal meaning from one context to another is to deny an explanation of much implied meaning. On the other hand, to argue that this necessarily happens always, just because it clearly often happens, is equally misleading.

An additional argument could be made that, quite simply, words like *alleviate*, *heal*, *relieve*, *soothe*, etc, have explicit/denotative/core evaluative meanings (see section 2.6). Or to use Whitsitt's own metaphor, it could be argued that these words are already semantically 'full' and that their meanings account for the fact that they reverse the evaluative polarity of their collocates. As we have seen, words that are said to have positive-negative semantic prosodies "are associated with polarity but might not be identified as evaluative out of context" (Hunston, 2011, p. 57), and this is precisely what Sinclair, Louw, Stubbs and others have found so compelling about semantic prosody in the first place. However, according the Collins Cobuild English Dictionary for Advanced Learners (CCED), "if you **alleviate** pain, suffering, or an unpleasant condition **you make it less intense or severe**" (CCED 2001, p. 40, emphasis added). Likewise, "If something **relieves** an unpleasant feeling or situation, **it makes it less unpleasant**

or causes it to disappear completely" (CCED 2001, p. 1303, emphasis added). Therefore, Whitsitt (2005, p. 296) is incorrect when he claims that "verbs like *alleviate*, *heal*, *relieve*, *soothe*, etc., [are] perfect candidates for semantic prosody since they all habitually appear in the company of clearly unpleasant words," since these the evaluative function of these words is explicit. Therefore, any attempt to assign a positive or negative semantic prosody to them, regardless of the evaluative polarity of their sets of collocates, is, quite simply, unnecessary.

1.3.2 Criticism two: semantic prosody is the same as connotation

Another of Whitsitt's more astute criticisms of semantic prosody that it is merely another term for connotational meaning. Whitsitt (2005, p. 285) argues that one of the more popular ways of defining semantic prosody "which is very widespread, treats semantic prosody as if it were a synonym of connotation." He cites Partington (1998, p. 66) as an example, but his apparent dismissal of the connection between semantic prosody and connotation is somewhat unfair to Partington (1998, p. 65), who argues that, in fact, "the term connotation is used to refer to at least three distinct phenomena." Partington equates semantic prosody with what he calls "expressive connotation" (as opposed to "social" or "cultural" connotation) but his argument is much more refined than Whitsitt gives him credit for. Far from arguing that semantic prosody is a synonym of connotation, Partington (1998, p. 66) devotes a chapter of *Patterns and Meanings* to "Connotation and Semantic Prosody" and is quite clear in his more nuanced position that semantic prosody is "one particularly subtle and interesting aspect of expressive connotation which can be highlighted by corpus data."

Morley and Partington (2009) also devote a section of their article "A few *Frequently Asked Questions* about semantic — or *evaluative* — prosody" to the question "Is semantic prosody

connotational?" The answer is that, like Partington, they view semantic prosody as an aspect of evaluative connotational meaning, but they make a key distinction, suggesting that "evaluative connotation is best considered as a cline" with items such as *murder* and *good* that clearly express (or 'denote') evaluative meaning on one end, and items such as SET *in*, HAPPEN, and CAUSE — the evaluative nature of which "were entirely obscure until assistance came to hand in the form of corpora" (Morley and Partington, 2009, p. 151) — on the other. A key point in this model is that "connotation is often considered to be more evident, less hidden, than semantic prosody" (Morley and Partington, 2009, p. 151). Putting semantic prosody on a cline of connotative meaning and considering it an aspect of connotation is subtly, but importantly, not the same as treating the two terms as synonyms, as Whitsitt maintains.

Similarly, in a summary of the phenomenon of semantic prosody, Hunston (2002, p. 142) writes: "It [semantic prosody] accounts for 'connotation': the sense that a word carries a meaning in addition to its 'real' meaning. The connotation is usually one of evaluation, that is, the semantic prosody is usually negative or, less frequently, positive." Arguing that semantic prosody "accounts for 'connotation'" does not necessarily mean that Hunston wishes to treat the two terms as synonyms, however. In the same summary Hunston (2002, p. 142) writes of semantic prosody:

> The semantic prosody of a word is often not accessible from a speaker's conscious knowledge. Few people, for example, would define *SET in* as meaning 'something bad starts to happen', but when the negative connotation is pointed out in many cases it accords with intuition (*A spell of fine weather set in* sounds very odd, for example).

This is arguably one of the key differences between the two terms, namely that an item's connotations can be accessed consciously while its semantic prosody is most often observable "only by looking at a large number of instances" (Hunston, 2002, p. 142) in a corpus.

Louw (2000, p. 50) is unyielding: "We need to make it plain that semantic prosodies are not merely connotational." He (2000, p. 51) argues that "[t]he force behind semantic prosodies is more strongly collocational than the schematic aspects of connotation." Louw cites the Collins Dictionary definition of *connotation*, quoted here from the CCED (2001, p. 317): "The **connotations** of a particular word or name are the ideas or qualities which it makes you think of." For example, under "connotation", the Literary Devices website[3] provides the example *home*[4], which "suggests family, comfort and security." However, the word form *home* (288,830) has only three positive collocates in the 6:6 Picture in the BoE[5]: *family* is 20th at N-1 with 2249 occurrences; *ideal* is forty-fifth at N-1 with and FaC of 887; and *win* is forty-fourth at N+1 with and FaC of 904. It is arguable, then, that a language user might consciously choose *home* instead of *house* in order to activate this positive connotation in the mind of a listener/reader, but this choice is not a collocational effect. That is, the positive feelings associated with home are not revealed by or in any way identifiable in corpus data.

Louw (2000, p. 51) argues that "when Philip Larkin describes *The Whitsun Weddings* as being '... like a happy funeral...', he is reversing a connotative pattern and not a semantic prosody." If we look at the word *funeral* in the Bank of English we see that it has only one obviously negative collocate, *death*, in the top-fifty list of collocates ranked by t-score. The only negative collocates in the 6:6 Frequency Picture for *funeral* are *died*, *death* (forty-sixth and fiftieth at N-6) and *killed* (forty-eighth at N+4, thirty-second at N+5, and forty-second at N+6). Not only is this very sparse collocational evidence of evaluation, the same picture contains the positive collocates *proper*, *decent*, *good* (all at N-1) and *romantic* (N+4).

To conclude, Partington (1998) and Morley and Partington (2009) may be correct that semantic prosody is a facet of connotational meaning, but it would not be correct to assume, as Whitsitt

does, that this makes semantic prosody and connotation synonymous, for the simple reason that "connotation can be collocational or non-collocational whereas semantic prosody can only be collocational" (McEnery, Xiao and Tono, 2006, p. 85). The collocational nature of semantic prosody sets it apart from more accessible (consciously) evaluative connotations.

1.4 Specific Objectives of the Thesis

Although I have dedicated a great deal of space in this chapter to defending positive-negative semantic prosody from criticism, it should not be assumed that semantic prosody is, therefore, unproblematic. The overarching argument of the thesis is that semantic prosody is potentially affected by at least four distinct linguistic phenomena.

First, Sinclair argues that semantic prosody is an obligatory category of the meaning of a lexical item, but collocation is an optional category. Most other models of positive-negative semantic prosody, however, claim that it is a collocational phenomenon (Louw, 1993, 2000; Stubbs, 1995; Bublitz, 1996; Partington, 2004; Adolphs, 2006; Xiao and McEnery, 2006; McEnery, Xiao and Tono, 2006; Bednarek, 2008; Morley and Partington, 2009; Walker, 2011a, 2011b; Barnbrook, Mason and Krishnamurthy, 2013). This leads to the first research question addressed in the thesis:

- What is the role of collocation in observations of positive-negative semantic prosody in corpus data?

This question is addressed by closely examining collocational profiles of the lemmas CAUSE and HAPPEN (see Section 5.2.1 for why these items were chosen). Discussion focusses first on the potentially problematic nature of statistical methods used to collect and examine evaluative collocates. Then analysis turns to the effects of phrasal collocates and finally considers the

relationship of collocates to the node word or phrase.

The second research question arose in response to comments made by Bublitz (1996), Sinclair (2003), and Partington (2004) about the phraseological nature of HAPPEN and its collocates. These researchers all make special note of the fact that HAPPEN is frequent in many semi-preconstructed phrases. Therefore, the research question is:

- How does phraseological behaviour affect observations of semantic prosody in corpus data?

This question is answered by looking at profiles and concordances of a number of phrases containing *happen*. This part of the investigation is split into two chapters (7 and 8), each of which focusses on different phraseological effects.

The final research question was inspired by the observation of a small sample concordance in Hunston (2007) of neutral instances of CAUSE. The concordance is constructed of lines taken from one register-specific sub-corpus of the BoE, and it was noticed that these lines evinced only two grammatical patterns of CAUSE. The research question, then, is:

- Is semantic prosody significantly affected by grammatical patterning, register, or both?

1.5 Organisation of the Thesis

This section outlines the organization of the thesis. Chapters 2, 3, and 4 provide the theoretical background of the linguistic factors that have been observed to affect the semantic prosodies of CAUSE and HAPPEN. Chapter 2 examines the notion of collocation as it relates to semantic prosody and establishes the parameters of collocation followed throughout the study. Chapter 3 looks at the phenomena of phraseology and grammatical patterning, specifically focussing on

how they are defined in the context of the current study. Chapter 4 examines the notions of register and genre in an attempt to disentangle them and establish a framework for analyses that follow. The next four chapters (Chapters 6 through 9) present the results of the corpus investigations into how these linguistic factors affect semantic prosody. Chapter 6 focusses on the potentially problematic nature of positive negative semantic prosody as an emergent collocational phenomenon. Chapters 7 and 8 look at the effects of phraseological behaviour on the semantic prosody of HAPPEN, and Chapter 9 discusses results of investigations into grammatical patterning and register. Chapter 10 discusses the pedagogical impact of the current findings. The final chapter concludes the thesis with a summary of the findings and suggests future avenues of research into semantic prosody.

Notes

[1] Although, if Louw is correct and semantic prosodies can be exploited for ironic effect, then they must be at least partially available consciously.

[2] Not to be confused with evaluative "core meaning" (Mahlberg, 2005, p. 149), which is discussed in detail in Section 2.6

[3] http://literarydevices.net

[4] https://literarydevices.net/connotation/ (accessed 26 October 2017).

[5] Please refer to Section 5.4.2 for a detailed explanation of the BoE collocational Picture output.

CHAPTER 2:
COLLOCATION

2.1 Introduction

Semantic prosody is widely considered a collocational phenomenon (Barnbrook, Mason, and Krishnamurthy 2013; Bednarek 2008; Bublitz 1996; Louw 1993, 2000; McEnery, Xiao, and Tono 2006; Morley and Partington 2009; Partington 2004; Stubbs 1995; Walker 2011a, 2011b; Xiao and McEnery 2006). However, defining an item's semantic prosody as an effect of its collocates, begs the theoretical question: what precisely constitutes a "collocate" and what is meant by the term "collocation"? This chapter explores some of the ways in which the terms collocate and collocation have been defined in order to establish how our understanding of these notions can affect observations of an item's semantic prosody.

Sinclair's model of the categories of co-selection, including collocation, which comprise the lexical item is discussed in Section 2.2. This is followed in Section 2.3 by a discussion of the notion of span and how it relates to calculations of collocational significance. In Section 2.4, various researchers' approaches to collocation are explored, and in Section 2.5 some specific approaches to semantic prosody as a collocational phenomenon are discussed in detail.

Finally, Section 2.6 discusses Mahlberg's continuum of evaluative meaning, of which semantic prosody is a key component. This leads to a detailed discussion of the necessity of observing close syntactic relationships between evaluative collocates and the node. Finally, potential difficulties presented by phrasal collocates are explored. The chapter concludes with an outline of how observations of semantic prosody are made in the analyses presented in Chapters 6 to 9 of this thesis.

2.2 The Lexical Item

This section introduces Sinclair's notion of the extended unit of meaning, "the lexical item" (Sinclair, 2004b), which comprises both collocation and semantic prosody in addition to three further categories of co-selection. Sinclair outlines the lexical item in two articles, "The search for units of meaning" (1996) and "The lexical item" (1998), both reprinted in the 2004 collection *Trust the Text*[1]. The impetus for creating this model of co-selection lies in the observation that "many, if not most, meanings require the presence of more than one word for their normal realization" (Sinclair, 2004b, p. 133). In the words of Hunston (2011, p. 55): "Units of meaning are identified by observing what commonly occurs in the co-text of a given word or short phrase and depend on identifying what is similar in a number of unique instances."

Sinclair (2004b, p. 141) characterizes the lexical item as follows:

> Five categories of co-selection are put forward as components of a lexical item; two of them are obligatory and three are optional. The obligatory categories are the core, which is invariable, and constitutes the evidence of the occurrence of the item as a whole, and the semantic prosody, which is the determiner of the meaning of the whole […]. The optional categories realize co-ordinated secondary choices within the item, fine-tuning the meaning and giving semantic cohesion to the text as a whole.

The lexical item's two mandatory categories of co-selection, the core and its semantic prosody, are presented by Sinclair as 'framing' the three optional categories — collocation, colligation, and semantic preference — which are in turn "related to each other in increasing abstraction; collocation is precisely located in the physical text," (Sinclair, 2004d, p. 142) and as such is the least abstract of the three. Colligation is observed when word classes are assigned to collocates and "where there is a preponderance of one particular word class" (Sinclair, 2004d, p. 142). Lastly, "[s]emantic preference requires us to notice similarity of meaning regardless of word class" (Sinclair, 2004d, p. 142). Sinclair does not claim that semantic prosody is abstracted from

semantic preferences, although arguments have been made that suggest that positive-negative prosody ought to be subsumed by semantic preference. For example, Bednarek (2008, p. 121) argues, "[t]he two types of collocation are […] very similar, differing only in degrees of 'generality', and frequently occur together," and Stewart (2010, p. 88) stops just short of arguing that "semantic prosody is primarily contingent upon semantic preference" out of deference to Sinclair's formulation of the lexical item.

It is especially significant to this thesis that, despite describing semantic prosody as the 'final' category of meaning (Sinclair, 2004b, 2004c), the pragmatic decision to express semantic prosody is in fact the initial choice made by the speaker: "The optional categories realize **co-ordinated secondary choices** within the item, fine tuning the meaning and giving semantic cohesion to the text as a whole" (Sinclair, 2004b, p. 141, emphasis added). Sinclair (2004c, p. 34), however, does "describe [the] elements in the unreversed sequence, the textual sequence." In this arrangement, "[t]he **initial** choice of semantic prosody is the functional choice which links meaning to purpose; all **subsequent choices** within the lexical item relate back to prosody" (2004c, p. 34 emphasis added).

This can be somewhat problematic because corpus-based studies intended to reveal an item's semantic prosody must necessarily work in precisely the opposite direction. That is, first a *core* is chosen for analysis and the corpus data — in the form of lists of collocates, positional frequency tables, or concordances — reveal significant collocations, colligational tendencies, and semantic preferences. Finally a decision is made as to whether these data can be taken as evidence of the "attitudinal" and "pragmatic" purposes of the speaker/writer, and "[h]aving arrived at the semantic prosody, we have probably come close to the boundary of the lexical item" (Sinclair, 2004c, p. 34).

2.3 Span

An important aspect of operationalizing the notion of collocation is the notion of span. The authors of The OSTI Report[2] (Sinclair, Jones and Daley, 2004, p. 5) write: "Collocation, or significant co-occurrence of lexical units, assumes that the extent of the environment, the 'co-', can be specified". Clear (1993, p. 276) uses a span of 2:2, that is two words on either side of the node, and Stubbs (1995) uses a window of 3:3 in his study showing the negative semantic prosody of CAUSE. Walker (2011a, 2011b), cites a 1974 study by Sinclair and Jones that recommends a span of 4:4 on statistical grounds. Stubbs (2001c, p. 29), citing the same paper observes that "[t]here is some consensus, but no total agreement, that significant collocates are usually found within a span of 4:4." McEnery and Hardie (2012, p. 129) argue, however, that in fact, "the majority of corpus linguists working on English have adopted Sinclair's guideline of a span of +/– 4".

Table 2.1 shows five concordance lines of the word form *cause* in the BoE that have been truncated to show only a span of four words on either side of the node.

Table 2.1: Example of the 4:4 span for five lines of CAUSE in the BoE

N-4	N-3	N-2	N-1	Node	N+1	N+2	N+3	N+4
to	therapy	because	they	cause	pain	to	others	and
cold	water,	it	can	cause	severe	burns	like	this
the	stupid	moods	that	cause	the	fall	out.	<h>
or	that	it	would	cause	them	to	be	shunned
because	It	does	not	cause	rashes.	Erma	says	that

Table 2.1 also shows the frequently used labels (N-4, N-3, etc.) for each position in the span. We can make use of these labels to show, for example, that the negative collocate *pain* is found at N+1 (one position to the right of the node) in the first line, while in the third line 3 the negative collocate *stupid* is found at N-3 (three positions to the left of the node).

2.3.1 Optimum span size

The writers of The OSTI Report (Sinclair, Jones and Daley, 2004, p. 42) ask "what in general is the range of influence of a node, that is to say, how many words away from a node must one go before the collocate there ceases to be affected by the node?" While it is technically true that there is no end to the range of influence a node might have on its collocates — The OSTI Report notes that "each node has an infinite region of influence, the influence decreasing the further away from the node you go" (Sinclair, Jones and Daley, 2004, p. 48) — it is possible to determine statistically the point at which the influence becomes negligible. The ultimate conclusion is that approximately 95% of significant collocates are found within the four-word span (Sinclair, Jones and Daley, 2004, p. 48). Figure 2.1 is the graph used in The OSTI Report (2004, p. 49) to illustrate the rapidly declining numbers of significant collocates[3] as the span increases.

Figure 2.1 Graph showing the "average node predictions over span positions 1-10", taken from Sinclair *et al.* (2004, p. 49)

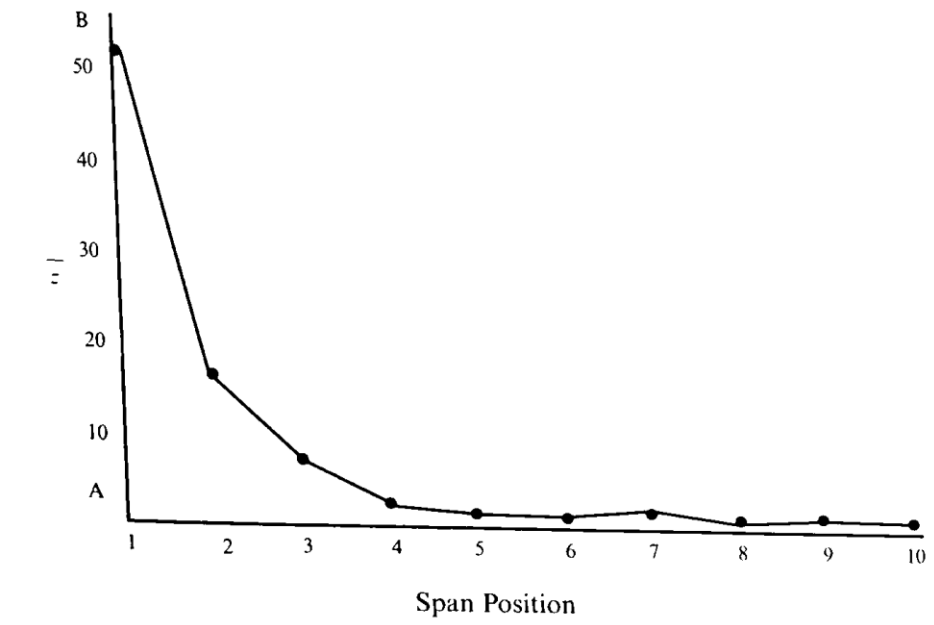


Figure 2.1 shows that the greatest influence (the highest number of significant collocates) is found at position one, followed by position two, and so on, and even at position ten, the influence of the node on its collocates does not reach zero. In an extreme case, "it has been claimed that significant statistical association can indeed be detected between a node and collocate which are separated by as many as fifty intervening words" (Clear, 1993, p. 276), but as Stubbs (1995, p. 8) argues, "this seems to alter the meaning of collocation, since the same content words are bound to occur at various points in a cohesive text." Mason (2006, p. 138) seems to agree, writing, "[t]here is no upper limit for the span, though it quickly becomes pointless once the value is too large: at a certain point the influence the node word has on its environment is overshadowed by the influences of other words." Indeed, one of the significant factors influencing the use of a window of four words is that at "about span position five onwards […] the number of significant collocates is the same as the number of mistakes one would make at the significance level used" (Sinclair, Jones and Daley, 2004, p. 43).

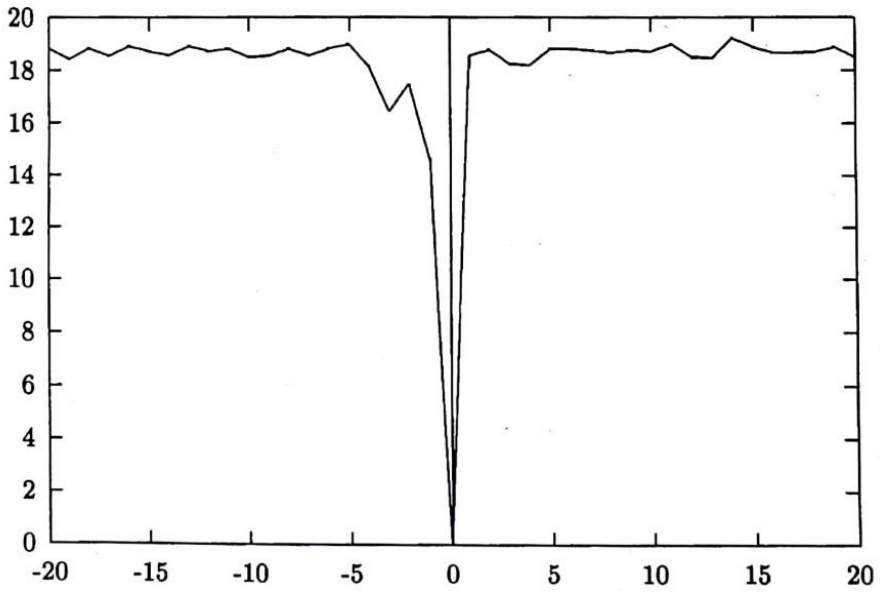
Both the British National Corpus (at bncweb.lanc.ac.uk) and Sketch Engine (at the.sketchengine.co.uk) default to a 5:5 span, but both allow the user to select other sizes. Interestingly, however, the BNC allows a minimum span of no less than 4:4 and a maximum of 10:10. Sketch Engine is more robust in that it allows for minimum spans of 1:0 and 0:1, and does not appear to have a built-in maximum (I tried 50:50 for the sake of testing the software and did not encounter any difficulties). The BoE List software (see Section 5.4.1) automatically calculates the FaC, t-score, or MI score for words that occur in a span of 4:4 and the span is not adjustable[4].

Mason (2000, p. 270) claims that the "basic assumptions" in previous studies, namely Berry-Rogghe (1973) and Sinclair and Jones (1974), "are not justified." He bases this criticism on the

results of Mason (1997) which concludes that each word has its own optimum span of influence and that these optimum spans are not necessarily symmetrical, as claimed in The OSTI Report. However, none of the data in either Mason (1997) or Mason (2000) suggests that there is good reason in the current study to look beyond the 4:4 or 5:5 spans that have become standard. In fact, not only does the variability in span size uncovered by Mason seem to be strongest for closed-class words (e.g. prepositions), the variation seems to appear in smaller span sizes rather than larger.

Figure 2.2 below is a graph taken from Mason (2000, p. 271) showing that the optimal span for the word *of* is the asymmetrical window of 5:1. On the left of the node the influence appears to level off at about position N-5, whereas on the right the node's influence does not extend beyond position N+1[5].

Figure 2.2 Graph showing the "The lexical gravity of *of*" in Mason (2000, p .271)



Comparisons of other graphs presented in Mason (2000) indicate that items have varying optimal span sizes, but none appear to extend beyond five words on either side of the node.

2.3.2 Span and semantic prosody

Stubbs (1995, p. 4, 2001c, p. 45) observes that the majority of evaluative collocates of CAUSE are found in a 3:3 or 4:4 span, but with respect to the semantic prosody of *lavish* he claims that "a 4:4 span is not always large enough to provide evidence of speaker attitude. Some collocates are embedded in longer sequences" (Stubbs, 2001c, p. 106). Similarly, Bublitz (1996, p. 15) explains that because of the complex structural nature of the collocates (see Section 2.6.2 for a detailed discussion) he uses a span of 15:0 in one of his investigations of HAPPEN. In the current study, the 4:4 span is utilized to establish an initial point of comparison, although in many cases, much larger spans are examined. These varying span sizes are primarily employed to allow for observation of antecedents of pronoun subjects and referents of general nouns, as well as structurally complex collocates (*cf*. Bublitz's 15:0 span used to study HAPPEN).

A final consideration is the fact that more precise locations of collocates may be more informative than the simple observation of their occurrence within a 4:4 window. That is, precise positional frequencies are likely to provide more information about collocational behaviour. It has been shown that even small differences in span size can have significant effects on the collocates observed. For example, McEnery and Hardie (2012, p. 128) show that when the span is increased from 3:3 to 5:5, only two of the top ten collocates of *cheese* in the British National Corpus (BNC) are the same. For this reason, corpus data that follow are most often taken from BoE Pictures, bespoke positional frequency tables, or observed directly in concordance lines (see Section 5.4 for detailed discussion terminology and methods used). Where Lists (see Section 5.4.1) are presented in the thesis, it is for general illustrative purposes, but more specific arguments employ Pictures, PFTs, and concordances.

2.4 Defining Collocation in the Literature

This section examines various ways in which the terms ‘collocation’ and ‘collocate’ have been defined in the literature. McEnery and Hardie (2012, p. 123) warn that “as soon as we […] attempt to pin down collocation either operationally or conceptually, we find a great multitude of different definitions.” Stewart (2010, p. 88) similarly notes that in the literature, “it is simply not the case that ‘collocates’ as noun [sic] always corresponds to ‘mere co-occurrences’, or, for that matter, that ‘collocation’ infallibly denotes habitual co-occurrence”. The discussion that follows is structured on Partington’s (1998, pp. 15–16) three classifications of collocation — also discussed in detail by Hoey (2005) — beginning with “textual” collocation in 2.4.1, followed by “psychological” or “associative” collocation in 2.4.2, and ending with a detailed discussion of statistical notions of collocational significance in 2.4.3. The section continues by discussing, in Section 2.4.4, potential difficulties applying the statistical model to the notion of semantic prosody, and concludes by suggesting in Section 2.4.5 an alternative statistical approach that can be applied to sets of low-frequency collocates.

2.4.1 Textual collocation: the co-occurrence model

Partington (1998, p. 15) begins his “definitions of collocation” by citing Sinclair as a primary proponent of “textual collocation”, quoting Sinclair (1991, p. 170) himself in support of the claim: “Collocation is the occurrence of two or more words **within a short space of each other** in a text” (emphasis added). Partington (1998, p. 15) paraphrases this notion as, “[o]ne item collocates with another if it appears **somewhere near it** in a given text” (emphasis added). However, Partington does not mention that elsewhere Sinclair is much more specific, taking pains to explain that the “short space” is, in fact, a “specified” amount of space, i.e. the defined span. For example, the authors of the OSTI Report (Sinclair, Jones and Daley, 2004, p. 10)

write: "[A] collocate is any one of the items which appears with the node **within the specified span**" (emphasis added), and in *Corpus Concordance Collocation* Sinclair (1991, p. 115) defines the term almost identically, using "the term *collocate* for any word that occurs **in the specified environment** of a node" (emphasis added). This notion of a "specified environment" or "span" is central to calculations of a collocate's statistical significance (discussed in detail in Section 2.4.3 below), and so it is somewhat misleading to argue that Sinclair advocates only for a simple co-occurrence model of collocation.

However, McEnery and Hardie (2012, p. 125) also argue that Sinclair, at least in his earlier research, "introduces an impressionistic approach to identifying collocation, based on manually scanning through the concordance lines." McEnery and Hardie (2012, p. 126) call this textual method "collocation via concordance." They point out that such an approach is common, citing "twenty papers in a festschrift for Sinclair" in which eight of eleven papers on collocation "use no statistical tests." Hoey (2005, p. 3), however, stresses that the textual definition "does not reflect Sinclair's own use of the term" and essentially dismisses the textual definition of collocation, arguing that "is not useful and can result in a woolly confusion of single instances of co-occurrence with repeated patterns of co-occurrence." Hoey does, however, recognise that reference to textual collocation is necessary at times; he mentions that instead of using the term 'textual collocation', "[w]henever I need to refer to the occurrence of two or more words within a short space of each other, I shall talk of 'lexical co-occurrence.'"

2.4.2 Psychological/associative collocation

Partington quotes Leech as a proponent of what he calls "psychological" or "associative" collocation: "Collocative meaning consists of the associations a word acquires on account of

the meanings of words which tend to occur in its environment" (Leech 1974, p. 20; as cited in Partington 1998:16). Partington (1998, p. 16) elaborates in his own words: "It is part of a native speaker's communication competence […] to know what are normal and what are unusual collocations in given circumstances."

Hoey appears to find the psychological approach to collocation the most useful but appends to it the importance of a statistical element, because the evidence for collocation is necessarily found in statistical analyses of corpus data. Hoey (2005, p. 5) writes:

> So our definition of collocation is that it is a psychological association between words (rather than lemmas) up to four words apart and is evidenced by their occurrence together in corpora more often than is explicable in terms of random distribution.

2.4.3 Statistical collocation

Partington quotes Hoey on statistical collocation: "'[C]ollocation' has long been the name given to the relationship a lexical item has with items that appear with greater than random probability in its (textual) context" (Hoey 1991, p. 6–7, as cited in Partington 1998, p. 16). As we have seen, Hoey (2005) himself does not ascribe solely to this definition, instead favouring a view that prioritizes the psychological/associative nature of collocation, while still recognizing the importance of statistical measures in identifying collocates in a corpus.

The statistical approach is also taken by Stubbs (1995, p. 1) who writes "[b]y collocation I mean a relationship of habitual co-occurrence between words (lemmas or word-forms)." Here, the "habitual" nature of collocation is quantifiable by the statistical methods Stubbs explores; indeed, one of the primary goals of this important article is to establish a stronger theoretical basis for statistical measures often employed in studies of collocation. To complicate matters somewhat, the collocation-via-concordance method is also attributed to Stubbs by McEnery

and Hardie (2012, p. 126), who point out that Stubbs argues "against the use of statistical significance calculations." In fact, these arguments against statistical calculation (see the following section for more on Stubbs' criticisms) might be seen simply as caveats that "statistics are not everything" (Stubbs, 1995, p. 14). Much later, Stubbs (2001c, p. 29) appears unequivocal in his approach to collocation:

> A 'node' is the word-form or lemma being investigated. A 'collocate' is a word-form or lemma which co-occurs with a node in a corpus. Usually it is frequent co-occurrences which are of interest, and Corpus Linguistics is based on the assumption that events which are frequent are significant. My definition is therefore a statistical one: 'collocation' is frequent co-occurrence.

A somewhat similar view appears to be adopted by Hunston (2002, p. 12), who writes "collocation is the statistical tendency of words to co-occur," and O'Keefe, McCarthy, and Carter (2007, p. 14) who define collocates as "word(s) [that] occur most frequently and with statistical significance (i.e. not just by random occurrence) in the word's environment." Tognini-Bonelli seems to agree but complicates the matter somewhat by introducing the term 'co-selection' and contrasting it with 'collocation': "The examination of these instances [of *fork out*] leads us to consider the issue of *co-selection*, that is the habitual selection of two or more items together, beyond the simple patterns of collocation seen above" (Tognini-Bonelli, 2002, p. 77, original emphasis). In this passage, 'co-selection' appears to refer to significant/statistical collocation, and 'simple patterns of collocation' appears to refer to the more general, textual, sense of collocation.

Although Partington has attributed the textual approach to Sinclair, Walker (2011a, p. 293) writes "[f]requency-based approaches are often associated with the work of Sinclair." This view is shared by Clear (1993, p. 277) who also attributes the statistical approach to Sinclair, writing "some linguists may consider every co-occurrence to constitute a *de facto* collocation. Sinclair's

work in this field, and my own, defines collocation as a **recurrent co-occurrence of words**" (emphasis added). In fact, The OSTI Report (Sinclair, Jones and Daley, 2004, p. 10) carefully disambiguates between textual collocation — "the co-occurrence of two items in a text within a specified environment" — and statistical collocation: "[s]ignificant collocation is regular collocation between two items, such that they co-occur more often than their respective frequencies, and the length of text in which they appear, would predict." From a broadly textual perspective, every word within the specified span is a 'collocate' regardless of frequency of co-occurrence or how it scores on significance tests. It is only when the statistical values calculated for these words exceed levels expected of random co-occurrences that the collocate is considered 'significant'. So it is true that Sinclair (1991, p. 170) espouses a textual definition, but he also notes that there is a "second kind of collocation, often related to measures of statistical significance" and that this second kind "is the one that is usually meant in linguistic discussions."

2.4.4 Difficulties with the statistical model of collocation

Barnbrook (1996, p. 94) writes: "The idea of significance […] relates directly to the concept of probability. In simple terms, a result is statistically significant if the probability of its chance occurrence is sufficiently low." The appropriateness of a wide variety of statistical measures of collocational significance has been discussed in great detail by a number of researchers. For example, the OSTI Report (Sinclair, Jones and Daley, 2004) discusses applications of the Chi-square, Fischer's Exact, and Poisson tests; McEnery, Xiao and Tono (2006) devote a section of their book to statistical measures of significance in which they discuss these same three tests as well as Z-score, log-likelihood, MI, and T-score tests; similar discussions can be found in Barnbrook (1996), Barnbrook, Mason and Krishnamurthy (2013), Hunston (2002), and elsewhere.

It is not possible here to discuss in detail these many diverse mathematical methods used to attach numerical significance to the relationship between node and collocate, but a brief explanation of the three methods pre-programmed into the Bank of English Lookup software — the main corpus used in the current investigation — will help to illustrate why reliance on statistical collocation is potentially problematic in studies of semantic prosody. It is important to note, of course, that more modern corpora interrogation software is often somewhat more robust than the BoE in the pre-programmed statistical measures they allow users to employ. The BNC, for example, allows for automatic calculation of Log-likelihood (the default setting), MI3, Z-score, T-score, Dice Coefficient, and of course, raw Frequency. Similarly, the enTenTen13 held by Sketch Engine allows for tests of T-score, MI, MI3, Log likelihood, Min. sensitivity, logDice, MI.log_f, and frequency.

The first and most direct method of observing collocational significance in the BoE is raw frequency, but, as McEnery et al. (2006, p. 215) argue, "[r]aw frequency is a poor guide to collocation." They show that it is difficult to judge significance based on frequency primarily because many co-occurring words are, in themselves, quite frequent and so are more likely to appear in frequency lists. For example, the most frequent collocate in the BoE List for CAUSE tagged as a verb is *the*. It is counterintuitive (at least) to consider *the* a collocate of any verb, and it is unlikely that *the* appears in this position because of a semantic or syntactic association with CAUSE. Its position is in part the result of being the most frequent word in the corpus, but also because it is frequently associated with the grammatical subjects and objects of the verb CAUSE. It will be argued in Section 2.6.1 that the syntactic relationship between collocate and node is an important consideration in observations of semantic prosody, and that simply occurring within the observed span, even at statistically significant frequencies, is an

insufficient condition for collocation, especially as evidence of semantic prosody.

A related method of establishing significance is a simple calculation that compares a word's observed frequency as a collocate of the node to its expected frequency of co-occurrence (Sinclair, 1991, pp. 69–70; Stubbs, 1995, p. 8; Barnbrook, 1996, p. 93; McEnery, Xiao and Tono, 2006, p. 215; Barnbrook, Mason and Krishnamurthy, 2013, pp. 60–64). It is simply a matter of calculating the statistical likelihood of one word following another in the corpus, where the raw frequencies of both (either within a span or at a specific location) and the size of the corpus are known. For example, we would expect to find CAUSE (89,830) *problems* (108,336) twenty-one times in the *ca.* 450-million-word BoE. This is calculated by first establishing the probability of encountering CAUSE in the 450-million-word Bank of English:

$$\frac{89{,}830}{450{,}000{,}000} = 0.0001996$$

This means that almost 0.02% of the corpus consists of CAUSE, or there are *ca*. 200 occurrences of CAUSE per million words. Thus, we expect to see CAUSE once in every 5,000 words. Similarly, the probability of *problems* occurring in the corpus is calculated as:

$$\frac{108{,}336}{450{,}000{,}000} = 0.0002408$$

Therefore, we would expect to see *problems* once every 4,153 words. Multiplying these probabilities together gives us the probability of *problems* following CAUSE in the BoE (or indeed, of CAUSE following *problems*; the equation does not account for the order of occurrence):

$$0.0001996 \text{ X } 0.0002408 = 0.00000004807$$

This means that the probability of encountering CAUSE *problems* (or *problems* CAUSE) in the BoE is one in almost 21-million words, so its expected frequency is just over twenty-one total occurrences. In fact, CAUSE *problems* occurs 1,488 times in the BoE, or almost sixty-nine times more frequently than expected by the calculation of random co-occurrence.

Sinclair (1991, p. 70) critiques this method of calculating statistical significance with an important observation: "The assumption behind this calculation is that the words are distributed at random in a text. It is obvious to a linguist that this is not so." This means that the fundamental principle underlying this method of establishing significance is the imagined state of a corpus composed of arbitrarily ordered words. That is, the calculation tells us, for example, that if the words of the corpus were randomly arranged, CAUSE would be expected once every 5,000 words. But language is never random. Stubbs (1995, p. 7) concurs with Sinclair, noting, "since textual data are never in this [random] form, this calls into question whether such statistics can reasonably be used on language data."

Stubbs (1995, p. 8) adds another important observation: "A problem with this calculation […] is that almost any observed co-occurrence is hundreds of times more likely than by chance." Stubbs gives an example of two words occurring 100 times each in a 1-million-word corpus and co-occurring only once. The expected frequency of this co-occurrence is only 0.01 times in the corpus, so the single co-occurrence is 100 times more frequent than expected by random association. Stubbs (1995, p. 8) contends: "But by definition, a single occurrence could just be due to chance. Such probability figures are artificially low, given that the data cannot be random."

The second statistical measure provided by the BoE is Mutual Information (MI) score. MI

scores are based on the observed-expected (O/E) calculation shown above, but now the result of O/E is "converted to a base-2 logarithm" (Hunston, 2002, p. 70). Barnbrook (1996, p. 98) provides the following formula:

$$MI = \log_2 \frac{O}{E}$$

Therefore, the MI score of CAUSE *problems* (1,488) is calculated as:

$$MI = \log_2 \frac{1{,}488}{21} = \log_2 70.86 = 6.15$$

The MI score is said to be "a measure of the strength of association between two words" (Walker, 2008, p. 80). An MI higher than 3.0 is generally considered significant, so the MI of *problems* (6.15^6) at N+1 of CAUSE in the BoE, indicates that the association between CAUSE and *problems* is moderately strong.

Corpus data shows, however, that MI rankings often provide lists of words lacking any clear semantic or syntactic association with the node. The primary difficulty with MI, argue McEnery et al. (2006, p. 217), is that it "gives too much weight to rare words." This appears to be especially problematic when observing collocates of high-frequency nodes in large corpora. For example, among the top twenty collocates of CAUSE ranked by MI, are *chagas* (4), *elephantiasis* (3), and *onchocerciasis* (3) (See Chapter 6 for more examples of such collocates).

The third statistical measure built in to the BoE software, the t-score, is also closely related to observed and expected frequencies, but in this calculation the expected frequency is subtracted from the observed frequency and the result is divided by the standard deviation (Hunston, 2002, p. 70). Lists ranked by t-score tend to be populated with "frequent words (whether or not they

are grammatical words) that collocate with a variety of items," (Hunston, 2002, p. 74). Barnbrook (1996, p. 97) provides the following simplified formula, arguing "the usual calculation of the standard deviation is considered to be unnecessary, and a useful approximation is used instead":

$$t = \frac{O - E}{\sqrt{O}}$$

The t-score of *problems* at N+1 of CAUSE then, is:

$$t = \frac{1{,}488 - 21}{\sqrt{1{,}488}} = \frac{1{,}467}{38.57} = 38.03$$

Whereas MI display the relative strength of association between two words, t-scores display level of "confidence that the association between [node] and [collocate] is genuine" (Stubbs, 1995, p. 11). A t-score higher than 2.0 is usually considered indicative of significance. Therefore, the t-score of *problems*, 38.03, indicates a very high degree of certainty that the association between *problems* and CAUSE is not the result of random occurrence. Again, since language is never random it is difficult to defend the t-score as much more than a very general indication of the relative significance of collocates. In the preface to the OSTI Report, Sinclair (2004, p. xxi) notes: "I may still use t-score for my day-to-day research in the absence of anything more plausible, but I have lost most of my original confidence in it and in other statistical procedures." Barnbrook et al. (2013, p. 89) agree, contending that, "[i]n fact, most statistical tests are not really applicable to linguistic data, as they assume a normal distribution, when word frequency counts are actually dominated by a few very frequent words followed by a large number of rare events."

Hoey (2005, pp. 3–4) makes a quite different argument against the use of such statistical

measures by pointing out that the statistical definition on its own "confuses method with goal," and "it gives no clue as to why collocation should exist in the first place." This is certainly the case in studies of semantic prosody where the semantic relationship between node and collocate is of primary significance, and it is difficult to sustain an argument that statistical significance is relevant. It may be true, as McEnery et al. (2006, p. 82) argue, that "the statistical approach to collocation is accepted by many corpus linguists […] in that they argue that collocation refers to the characteristic co-occurrence of patterns of words," but as Stewart (2010, p. 86) observes, "the application of this [statistical models of collocational significance] within the domain of semantic prosody is not systematic." Stewart (2010, p. 86) notes that neither Louw (1993) nor Bublitz (1996), for example, appear to require statistical significance of co-occurring words for them to be observed to evince an item's semantic prosody. Further, Stewart quotes Hoey (1997, p. 5), who argues[7]:

> When a new disease is found, it can immediately be added, for example, to the list of things that can be *caused*; we do not have to wait until it has become common enough for it to figure in calculations of collocations.

2.4.5 Total FaC values as indicators of strength of semantic prosody

In Section 2.2 the causal relationship between semantic prosody and the three optional categories of meaning was explored. Specifically, it was emphasized that semantic prosody is the initial pragmatic selection and that the optional categories are secondary choices, essentially 'filling out' the meaning of the lexical item. This clarification is critical to analyses that follow in the thesis.

The most salient objection to requiring statistical significance of collocates evincing semantic prosody is quite simply that semantic prosodies are observed in groups of collocates that share the semantic property of evaluating positively or negatively. The significance (or otherwise) of

a single collocate is not relevant to observations of semantic prosody. McEnery and Hardie (2012, p. 136) are the only writers encountered in the preparation of this thesis who are explicit about this issue. They write:

> The negative things are not, necessarily, themselves significant collocates […]; it is when they are considered in the aggregate that their frequency becomes notable. Thus, an analysis of semantic prosody is an abstraction across multiple, different contexts of usage.

Stubbs (1995, p. 14) accounts for low-frequency, semantically relevant collocates by suggesting that they be included in statistical profiles by calculating a single t-score for the group. He argues that this single t-score displays the degree of confidence that the association between the group and the node is genuine, in the same way that it does for a single word. The analyses presented in this thesis do not follow Stubbs' recommendation directly (the calculation of a combined t-score is considered an unnecessary extra step). A simpler procedure is employed throughout analyses that follow although the general intention is the same.

The combined raw Frequency as Collocate (FaC) of groups of evaluative collocates is informative, especially when observation of both positive and negative collocates creates potential ambiguity regarding the polarity of item's semantic prosody. For example, corpus data presented in collocational profiles in Chapters 6 to 9 often show apparent conflicts between the number of evaluative collocates and their total FaCs. Table 2.2 shows a hypothetical data set to illustrate this point. As the table shows, we might imagine a node that occurs in the corpus a total of thirty times and has fifteen unique collocates. If ten of these collocates are positive and five are negative, we would likely argue for a positive semantic prosody.

Table 2.2 shows, though, that if we consider the summed FaCs of the evaluative collocates, the node appears to have a negative prosody. In the imagined example, five collocates account for

twenty occurrences, or 66.6% of the thirty lines. Because semantic prosody is the initial functional/pragmatic choice, it is clear that speakers activate the negative prosody more frequently even though this prosody is activated by a smaller number of collocates.

Table 2.2: Illustration that total evaluative FaC scores are often more salient than numbers of evaluative collocates in judgments of semantic prosody

Positive	FaC	Negative	FaC
Collocate 1	1	Collocate 1	4
Collocate 2	1	Collocate 2	4
Collocate 3	1	Collocate 3	4
Collocate 4	1	Collocate 4	4
Collocate 5	1	Collocate 5	4
Collocate 6	1		
Collocate 7	1		
Collocate 8	1		
Collocate 9	1		
Collocate 10	1		
	10	Total	20

This method is employed throughout the analyses that follow in this thesis and is highlighted in a number of cases where disambiguation is required.

2.5 Semantic Prosody as a Collocational Phenomenon

This section discusses the notion of collocation as it relates specifically to semantic prosody. McEnery and Hardie (2012, p. 136) explicitly appear to take for granted that semantic prosody "is a concept rooted in the neo-Firthian **concordance-based analysis of collocation**" (emphasis added), and instead of describing semantic prosody in terms of collocates or collocation they write: "Words or phrases are said to have a negative or positive semantic prosody if they **typically co-occur with units** that have a negative or positive meaning" (emphasis added).

Likewise, nowhere does Sinclair refer to semantic prosody as a 'collocational' phenomenon. In fact, in Sinclair's model semantic prosody simply cannot be strictly collocational because

semantic prosody is a required element of co-selection and collocation is not; if semantic prosody were collocational (in a strong sense), then collocation would have to be a necessary category of co-selection. Task 14 of *Reading Concordances: An Introduction* (Sinclair, 2003, p. 117), in which Sinclair refers directly to "the semantic prosody of *happen*", contains no mention of 'collocation' or 'collocate(s)'. Instead, throughout the task, Sinclair (2003, p. 124) writes of co-occurring "events" and "expressions", as in the following summary of the behaviour of *happen*:

> The main orientation of *happen* is **the prospection of an unfortunate event** happening; this often **goes with expressions of doubt and vagueness**. Occasionally the word presages the opposite - a desirable event - and in such cases **there are often expressions of certainty along with it** [emphasis added].

Near the end of the task Sinclair (2003, p. 125) writes of words "occurring together" and "the notion of CO-SELECTION [sic]" which he defines as "the simultaneous choice of more than one word at a time," but still, neither "collocate" or "collocation" are used to describe this relationship.

Similarly, Susan Hunston does not use the terms 'collocate' or 'collocation' in any of her explications of, and references to, semantic prosody. Referring to Stubbs (1996:188), Hunston (2002, p. 119) writes, "the word *intellectual* **co-occurs with words** which many people would regard as negative" (emphasis added). Hunston (2002, p. 141) also writes that semantic prosody "usually refers to a word that is typically used in **a particular environment**" (emphasis added), and as an example she notes that *SIT through* has a negative semantic prosody "[b]ecause it is often **used with items** that indicate something lengthy and boring" (emphasis added). On the same page Hunston cites an example by Louw: "[*I*]*n vain* is usually used **in the context of** something involving effort and intention" (emphasis added). In her influential article "Semantic

Prosody Revisited", Hunston (2007, p. 259) uses "collocation(s)" three times and "collocational" once, but all of these occur in a single paragraph explaining how "collocational inference" is used to explain "subjective reactions to individual instances of language." Throughout the remainder of this important article, no further mention of collocates or collocation is made.

However, many researchers do explicitly claim that semantic prosody is collocational. Stewart (2010) discusses the issue of collocation in depth in his book-length treatment of semantic prosody, and begins by summarizing what he considers "[t]he most common interpretation" of the phenomenon. Stewart (2010, p. 1) writes:

> Semantic prosody is instantiated when a word such a CAUSE co-occurs regularly with words that share a given meaning or meanings, and then acquires some of the meaning(s) of those words as a result. This acquired meaning is known as semantic prosody."

Louw (1993, p. 157) defines semantic prosody as "a consistent aura of meaning with which a form is imbued by its collocates." Bublitz (1996, p. 9) describes the phenomenon thus: "[t]he node itself is […] habitually associated with its semantic prosody, which is based on a semantically consistent set of collocates." Similar characterizations are found throughout the literature (Barnbrook, Mason, and Krishnamurthy 2013; Bednarek 2008; Bublitz 1996; Louw 1993, 2000; McEnery, Xiao, and Tono 2006; Morley and Partington 2009; Partington 2004; Stubbs 1995; Walker 2011a, 2011b; Xiao and McEnery 2006).

Adolphs (2006, p. 69), however, defines the relationship in a subtly different, and perhaps more accurate way: "The shading of a lexical item can be determined by looking at its collocates." Adolphs appears to be acknowledging that while an item's semantic prosody is often observed in the set of semantically related (i.e. evaluative) collocates, the prosody is not created by that set of collocates. As we have seen, the causal chain moves from the initial pragmatic/functional

choice of prosody to collocation. The apparent misunderstanding that semantic prosody arises in the opposite direction, i.e. from the process of collocation, is likely due to the extremely close associations among the categories of meaning comprising the lexical item. Sinclair (2004b, p. 142) explains:

> Semantic preference requires us to notice similarity of meaning regardless of word class; however there may well be found within a semantic class one or more colligations of words which share both the semantic feature and a word class. There may also be collocates, specific recurrent choices of word forms carrying the semantic preference.

Elsewhere Sinclair (2004c, p. 35) notes that "in a number of cases we find that the semantic preference and the semantic prosody are fused."

Indeed, semantic prosody and semantic preference are now often considered inseparable. Stewart (2010, p. 88), nearing the end of a detailed argument about the nature of collocation and co-selection, presents a stronger position: "It could with justification be argued that what semantic prosody is primarily contingent upon is semantic preference, and that whether the item has a relationship of *habitual* co-occurrence with any of its co-text is something of an irrelevance." Stewart does not fully commit to this perspective in deference to Sinclair's model in which the two phenomena are distinct.

Partington (2004, p. 151) makes the similar claim that at times "semantic preferences combine to form (or reflect) an overriding prosody." Partington (2004, p. 149) initially appears to suggest that we might subsume semantic prosody under the category of semantic preference:

> One view would be that semantic prosody is a sub-category, or a special case, of semantic preference, to be reserved for instances where an item shows a preference to co-occur with items that can be described as bad, unfavourable or unpleasant, or as good, favourable or pleasant. This description of the relationship between the two phenomena as set and sub-set, probably works as a rule of thumb in most cases.

However, Partington (2004, p. 150) continues by claiming that "the relationship [between prosody and preference] is more complex and the difference more fundamental than the above description suggests." So while Partington ultimately insists on maintaining a clear division between the two, he does affirm that "preference […] contributes powerfully to building […] prosody" (Partington, 2004, p. 151).

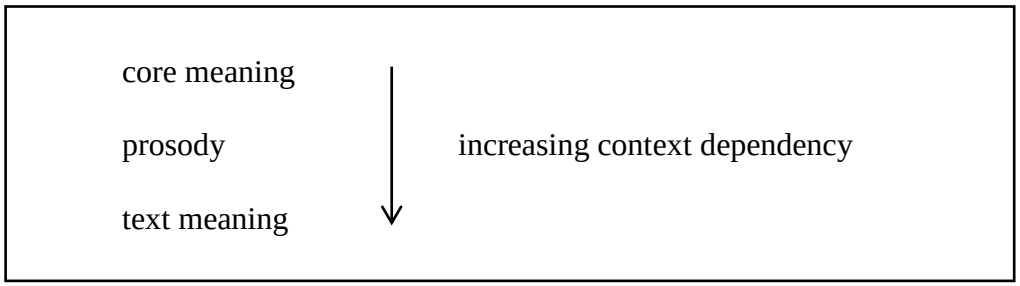
It remains problematic that semantic prosody is primarily observed via collocation. The salient question would seem to be, then, does a collocate have to be statistically significant to be evaluatively significant? Of course, frequency of a sort is a principal factor; the fact that a number of evaluative collocates are observed to have been co-selected with the node is precisely why we are able to say that the node has a semantic prosody in the first place, and as outlined above, the combined frequencies of the co-occurring evaluative words will come to bear on claims for or against positive-negative prosodies in the discussions that follow. However, the precise relationship between collocation and semantic prosody remains an area deserving of further research.

2.6 The Continuum of Evaluation

Mahlberg (2005), like Sinclair and Hunston, does not refer to collocates or collocation in her description of semantic prosody. Instead, Mahlberg (2005, p. 149) argues for a tripartite approach to observations of how the evaluative nature of lexical items are observed in corpus data. She argues that an item's core evaluative meaning, evaluative prosody, and textual evaluation "differ with regard to the type of context that is needed to describe them: from the core meaning to the text meaning, the evaluative facets of the meanings of words increase in their dependency on context to become discernible." Malhberg's continuum of evaluative

meaning is illustrated in Figure 2.3 below.

Figure 2.3 "Evaluative meanings of lexical items in context", taken from Mahlberg (2005, p. 149)



Core evaluation is revealed when speakers are aware of an item's evaluative meaning without requiring any additional context and when they consciously use the item for the purposes of evaluation. Examples of words displaying this type of evaluative meaning are *splendid*, *happily*, *triumph*, *win*, and *lose* (Mahlberg, 2005, p. 149). Further along the cline is evaluative prosody which requires more context to be realised, though precisely how much context (i.e. the optimum span) is not specified, and, indeed, is likely impossible to delineate in absolute terms. Unlike core meaning, this type of meaning is unavailable to intuition (Mahlberg, 2005, p. 150) and is revealed only through close analysis of large numbers of attested instances in corpora. Finally, evaluative text meaning in corpora requires the most context and "refers to uses of words that do not typically express evaluation, but depend to a larger extent on a specific text. Evaluative meaning as text meaning can include instances where we may not even find evaluation markers in the context" (Mahlberg, 2005, p. 150).

Following are examples from the BoE that illustrate the difference between prosody and text meaning. Line 1 shows, for example, an instance of evaluative prosody where there are no evaluative collocates, positive or negative, within the 4:4 span; *inefficiency* and *failed* are found

at N-7 and N-5 respectively, and *errors* is at N+5.

1. Mr Darling also blasted inefficiency and failed working practices that have **caused** the huge number of errors in administrating benefits.

Despite the fact that the collocates occur outside of the standard span of collocational influence, this is not an example of textual evaluation (Mahlberg, 2005, p. 150) because there is no need to apply additional contextual or linguistic knowledge to observe that that which is *caused* is negative. The evaluation provided by *inefficiency*, *failed*, and *errors* is collocational in the simple 'co-occurrence' or collocation-via-concordance sense outlined above (Section 2.4.1).

In line 2 there is one clear single-word evaluation marker, namely *blame* at N+10, but as evidence for semantic prosody this is not particularly strong.

2. Good idea," he said, and went to deliver a calf which had two heads. **These things happen**, with or without radiation. The farmer was inclined to blame Chernobyl, all the same, and

However, text evaluation is found in the clause *and went to deliver a calf which had two heads* which is the referent of the phrase *these things happen*. The co-selection at N+11 of Chernobyl also indicates textual evaluation. In addition, it is possible to argue that virtually anything that happens at or associated with Chernobyl can be considered unfavourable. This might be considered a borderline case, however, because it takes only very little contextual knowledge to surmise that the proposition expressed in line 2 is, in fact, negative.

In line 3, however, the evaluation provided by the phrase *make things happen* is clearly textual. The phrase is used to refer to a general, indeterminate, set of contextually relevant propositions. What is specifically expected of *his players* is not stated, and there are no explicit evaluative markers to be found.

3. its marker defence and kick/chase. Pearce will emphasise to his players they must **make things happen,** rather than wait for things to unfold. The Blues completed training yesterday with a run

The neutral collocates *defence*, *kick/chase*, *Pearce*, *players*, *Blues*, and *training* combine to create a context within which a reader with sufficient knowledge of the subject matter would understand the evaluative intention (positive) of the writer.

The notion of a cline of evaluation that depends on varying degrees of context for each mode's realisation is important for a number of reasons. First, it may help to explain why Stubbs (1995, p. 18) initially finds no evidence of the negative semantic prosody in negative collocates of HAPPEN: "the negative collocates are missed with a window of 4:4." Only when he expands his search to a span of 8:0 does he observe a predominantly bad set of collocates. For the same reason, Bublitz expands his search for negative collocate of HAPPEN to a window of 15:0, and then still finds "only some" evidence. Bublitz does not deny that HAPPEN has a negative semantic prosody, in part out of regard for Sinclair who made the original claim that the things that happen tend to be negative, but also because such a claim "is also clearly counter-intuitive" (Bublitz, 1996, p. 18). We might now argue that the reason Bublitz finds only some evidence of semantic prosody and why this might seem counter-intuitive is that the negativity expressed in the environment of HAPPEN is in fact textual and not prosodic at all.

2.6.1 Collocates' association with the node

Closely related to the concept of the amount of context required to observe the three modes of evaluation is the notion of the collocates' logical and syntactic relationship to the node. As Barnbrook et al. (2013, p. 164) argue:

> Part of the problem in using collocation to model language is that the set of words co-occurring with a particular node word is rather mixed. There are no restrictions on the

> word-class, or the position of the collocates relative to the node, or even the relationship between the node and the collocate.

Similarly, in an endnote Partington (2004, p. 154) makes an important "additional observation" and argues convincingly that, in studies of semantic prosody,

> the logical relationship of an item to its collocates is a vital consideration. Simply being primed to appear in the environment of collocates of a certain sense, good or bad, is not a sufficient condition for an item to acquire the same sense.

Partington is referring to criticisms of semantic prosody (see especially Whitsitt 2005) that it is not always the case that a word takes on the evaluative meaning of its collocates (see Section 1.3.1 above). Partington provides the examples of *relief* and *ease* which, he argues, are inherently positive yet collocate with "unfavourable words like *debt, pain, poverty, suffering* and so on. In the current study, Partington's claim that the logical relationship between node and collocate is critical, and this notion is extended here to mean more specifically that for a collocate to be considered evidence of semantic prosody, a clearly observable close syntactic relationship is required. This means that we cannot accept just any evaluative word in the context of the node (irrespective of the size of that context) as evidence of semantic prosody.

Lines 4 and 5 were selected to illustrate the point because in each case an evaluative collocate is found across a clause boundary. In 4, *terrible damage* is collocational evidence of negative semantic prosody, but *shame* is not. Likewise, *problem* and *failures* evince the semantic prosody of *cause*, but *difficulty* does not.

```
4. terrible damage these substances can cause. <p> The expressions of shame
5.   the problem that such failures can cause. However, one difficulty the
```

These lines are merely illustrative of syntactic relationships between collocate and node and do not challenge the claim that CAUSE has a negative semantic prosody.

In line 6 below, however, *confidence* and *best* suggest positive textual evaluation.

```
6. confidence, try my best and see what happens," she said. <p> Whoever is
```

Here, *see what happens* refers to an unknown future outcome. However, because the phrase expresses "doubt or uncertainty" (Sinclair, 2003, p. 117), we might argue instead that the line is textually negative. Regardless, since nothing has in fact happened, this instance is not considered to contain evidence of semantic prosody.

2.6.2 Phrasal collocates

Finally, another potential problem describing semantic prosody as a collocational phenomenon is the presence of phrasal collocates in corpus data. Evaluative phrases are potentially problematic primarily because they do not appear in Lists or Pictures. Stubbs (1995, pp. 15–16), for example, observes that some statistically significant collocates of CAUSE, in the context of a collocational profile, appear neutral or positive, but are in fact elements of phrases that evaluate negatively:

> The frequent collocation with *great* is partly due to phrases such as *cause for great concern*. Similarly, a frequent collocate of CAUSE is *driving*, not because the words directly collocate, but because of the phrase *reckless driving*, which in turn occurs in phrases such as *death caused by reckless driving*. Another collocate is *natural*: due to occurrences of *death from natural causes*.

Stubbs calls phrases like these a "problem", and notes that "[s]uch inter-collocations are beyond the scope of the methods discussed here" (Stubbs, 1995, pp. 15–16).

Barnbrook et al. (Barnbrook, Mason and Krishnamurthy, 2013, p. 172) describe a similar problem,

> that of syntactic dependency: *unfair* should not really be a collocate of *claim*, as what is

> usually *claimed* is *unfair dismissal*. But because this is such a common bigram, *unfair* and *dismissal* both show up independently as collocates of *claim*. What would ideally be needed to solve this problem is a filter which removes words that act as modifiers to other words from the data, making use of linguistic knowledge for improving the result.

Tognini-Bonelli (2001, p. 24) encounters similarly problematic phrases in her analysis of the semantic prosody of *flexible*. She concludes:

> When it comes to identifying the semantic prosody, the collocational profile is not much help […]. With *flexible* we find that the positive evaluation is realised in a variety of ways which are not picked up by a computer program that focuses on the recurrent coselection of individual words. A collocational profile is best read as a confirmation of observations in the concordance, after the analyst has familiarised him/herself with the repeated patterns".

Investigations that follow begin with quantitative analyses of collocational profiles, in spite of the fact that neither phrasal collocates nor the syntactic/semantic relationships between collocates and node can be observed in such data sets. The fact is most often the collocates in these profiles are found in close syntactic/semantic association with the node, and, in the case of positional frequency tables (PFTs) phrases can often be identified. Therefore, it is often possible to identify potential evaluative trends that can then be confirmed and defined in concordance analysis. Moreover, frequent substitutable items in semi-preconstructed phrases, calculation of total evaluative FaCs and making comparisons between groups of evaluative collocates, and so on, are only possible in collocational analysis. For these reasons, collocational profiles are central to examinations presented in Chapters 6-9. These profiles are in every case, however, followed by discussions of qualitative analyses, and, as Tognini-Bonelli suggests, the results are, at times, significantly different and worthy of detailed discussion.

2.7 Conclusion

On a purely theoretical level, characterizing positive-negative semantic prosody as a

collocational phenomenon may not be, strictly speaking, correct. It is perhaps more accurate to suggest that evaluative collocates, indirectly through semantic preference and colligation, are an effect of positive-negative semantic prosody. That is, evaluative collocates are textual realizations of the initial abstract functional decision to evaluate. However, semantic prosody is generally considered collocational not because there is an observable causal relationship between the two phenomena, but simply because prosody is, most often, abstracted from groups of collocates that evaluate similarly. This may seem an obvious point, but in the chapters that follow, it will prove to be critical.

On a more practical level, in studies of semantic prosody it would seem necessary to be maximally unambiguous in outlining what is meant by the terms "collocate" and "collocation" in order to avoid the potential problems discussed in this chapter. Specifically, four main aspects of collocation as evidence of semantic prosody should be defined whenever possible:

1. Does the term collocate refer only to single words or does the study include phrasal collocates?
2. Is statistical significance a requirement of collocates evincing semantic prosody? If so, which statistical measure is used to calculate this significance?
3. Explication of statistical measures necessarily also requires an explicit account of the span employed in such calculations. However, even if statistical significance is not a requirement, the span within which collocates are collected remains a key factor, as larger spans may indicate textual evaluation.
4. What, if any, relationship (syntactic, sematic, logical) between collocate and node is required for the collocate to be considered evidence of semantic prosody?

In this thesis, a collocate's significance is considered evaluative, rather than mathematical. Statistical significance may be a sufficient condition for collocates that evince semantic prosody, but it is not a necessary one, as will be shown in many of the analyses that follow. The term 'collocate' is, therefore, used for any word (or phrase) that occurs within a specified

environment (span, clause, sentence, number of characters) regardless of significance scores. Specific spans are made explicit in each study that follows in Chapters 6 through 9. Further, it will be demonstrated more than once in this thesis that strict adherence to standard notions of significance can severely limit the amount of available data, potentially making any claims for or against the semantic prosody of an item problematic, if not impossible. This issue is especially prominent in examination of the semantic prosody of longer phrases which occur in the corpus much less frequently than single words and therefore necessarily reveal fewer collocates for analysis.

However, an element of automation in data collection is necessary due to the vast amounts of data in consideration. The lemmas CAUSE and HAPPEN tagged as verbs occur 89,830 and 149,408 times in the BoE, which means that manual examination of 500-line concordances represents only 0.56% and 0.33% of the total occurrences respectively. Therefore, a method of automatic retrieval of high-frequency collocates was considered an essential part of the analysis. Therefore, despite the fact that "a relationship to chance is not likely to be very revealing" (Sinclair, 2004c, p. 29), T-Lists and T-Pictures are used frequently throughout the thesis primarily as a technique to "amass examples" (Stubbs, 1995, p. 14) (or, conversely, to observe the absence of examples).

To conclude, in this thesis, semantic prosody is argued to be activated when:

- The collocate — realized as a single word or phrase, regardless of the span within which it is observed — is clearly evaluative and in close syntactic association with the node. The semantics of the collocate are taken into account, i.e. in the cases of *cure disease*, *alleviate pain*, neither *disease* nor *alleviate* are considered evidence of negative semantic prosody (see Morley and Partington 2009:142).

- In addition, where the antecedent of a pronoun or referent of a general noun is clearly observable and evaluative, it is considered evidence of semantic prosody, with the caveat that, pragmatically, a span must be enforced; 200-characters was chosen arbitrarily for some of the analyses presented and even this large span may not be enough. However, spans of this size and larger make automatic and semi-automatic identification of evaluative collocates prohibitively challenging.

An additional point central to observations of semantic prosody in this thesis:

- The number of collocates is less relevant than the number of occurrences of the collocates (their summed FaCs). This is because the total occurrences indicate the number of times the initial pragmatic/functional decision to evaluate is made, regardless of which collocates is chosen to realize that decision.

Notes

[1] Citations of both articles refer to the 2004 publication.

[2] The original report was completed in early 1970, and is cited here from the 2004 publication.

[3] Measures of significance, in this case a variation of z-score, cut-off thresholds, etc. are explained in detail earlier in The OSTI Report.

[4] Though the Picture and Collocations programs allow the user to specify larger or smaller spans.

[5] Mason (1997, p. 101) explains the calculations underlying the notion of lexical gravity in more detail. The vertical axis shows the Type-Token Ratio (TTR) as a percentage. Therefore, in Figure 2.2, we would expect only about fourteen different words at position N-1 in a random sample of 100 instances of *of*. At N+1 we would expect about nineteen different words, and so on.

[6] The BoE, presumably using a slightly different calculation, reports the MI of problems at N+1 of cause as 6.0979.

[7] Hoey (2005, p. 23) makes specific reference to such claims and clarifies his position, writing that he "would ask readers of those papers to interpret [his] references to semantic prosody as references to semantic [preference]".

CHAPTER 3:
PHRASEOLOGY AND GRAMMATICAL PATTERNING

3.1 Introduction

As we have seen, semantic prosody is one of the categories of co-selection comprising the lexical item, and as such, "[t]he semantic prosody of a lexical item is a consequence of the more general observation that meaning can be said to belong to whole phrases rather than to single words." This chapter explores this relationship in more detail. First, Section 3.2 establishes the foundation of the discussion by outlining Sinclair's two general principles of meaning creation, the "open-choice" and "idiom" principles. Section 3.3 focusses on approaches to the problem of defining specific phraseological items. This section focuses on approaches by Wray (2002) and Gries (2008) because they both begin their respective studies by acknowledging the vast number of often disparate definitions and methods of classifying phraseological structures. This leads to a discussion in Section 3.4 of two approaches to phraseological research as outlined by Granger and Paquot (2008), who make a clear distinction between the deductive and inductive approaches to phraseology. The former involves placing known phraseological items on clines of opacity and fixedness, "with the most opaque and fixed ones at one end and the most transparent and variable ones at the other" (Granger and Paquot, 2008, p. 28). The latter is a statistical approach and involves generating phraseological items from corpus data. This leads to more detailed discussions in Section 3.5 of three ways in which statistical data can be applied in studies of phraseology: n-gram/cluster analysis; MWUs defined by cumulative frequency; and collocational frameworks. This discussion of frequency is then continued in Section 3.6, where the notion of defining a canonical variant of a semi-preconstructed phrase is explored. Finally, Section 3.7 turns to grammatical patterning and the relationship of patterning to meaning.

3.2 Sinclair's Open-choice and Idiom Principles

In an article titled "The phrase, the whole phrase, and nothing but the phrase" Sinclair writes "we have to concede that the normal primary carrier of meaning is the phrase and not the word" (2008, p. 409). Indeed, much of John Sinclair's research career centred on establishing that the phrase, not the single word, is the primary repository of meaning in language. Sinclair's chapter "The nature of the evidence" in 1987's *Looking Up: An account of the COBUILD Project in lexical computing* (much of which is reprised in 1991's *Corpus Concordance Collocation*), for example, is primarily concerned with the phrasal verb SET *in*. Here he (1987, p. 150) argues that one of the reasons such phrasal verbs can be problematic for teachers and learners is that "[t]he co-occurrence of two quite common little words can unexpectedly create a fairly subtle new meaning that cannot be systematically related to either or both of the original words." It is worth recalling, too, that this analysis of *SET in* contains perhaps the earliest formulation of what would later be known as semantic prosody (see Section 1.2).

In his 1991 book, *Corpus Concordance Collocation*, Sinclair (1991, p. 109) furthers the theoretical grounding of phraseology as the primary carrier of meaning in text, writing:

> It is contended here that in order to explain the way in which meaning arises from language text, we have to advance two different principles of interpretation. One is not enough. No single principle has been advanced which accounts for the evidence in a satisfactory way.

Sinclair (1991, p. 109) defines the first of the two, "the open-choice principle", as the "normal" way of accounting for meaning in text. It is the "slot-and-filler model" used by most grammars to explain how words come together to create meaning: "At each point where a unit is completed (a word, phrase, or clause), a large range of choice opens up and the only restraint is grammaticalness." The primary reason Sinclair posits the necessity of a second principle is that

"[c]omplete freedom of choice […] of a single word is rare." (Sinclair, 2004c, p. 29). One of the foundations of Sinclair's argument for the lexical item is that only very infrequently occurring words and specialized terminology are selected using the open-choice principle (although he suggests that even technical terms may not be as independent as they are often thought to be). Sinclair argues that more frequent words "retain traces of repeated events in their usage, and expectations of events such as collocation arise" (Sinclair, 2004c, p. 30).

Thus, in contrast to the open-choice principle, Sinclair puts forth "the idiom principle". Sinclair (1991, p. 109) writes: "The principle of idiom is that a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments." Sinclair's claim is that there are, in reality, a number of "restraints on consecutive choices" that extend beyond mere grammatical acceptability at each point in an utterance. Sinclair (1991, p. 110) cites some of the factors that affect how words come together in text: "the nature of the world around us" (words referring to physical objects found together in the world are likely to be found in texts together: no examples are provided but they are easy enough to imagine, e.g. *table* and *chair*, *teacher* and *student*, etc.); related philosophical concepts (again, many examples spring to mind, like *love* and *marriage*, *war* and *peace*, etc.). He also cites organizational principles "such as contrasts or series." Finally, he argues that register likely has an effect on phraseology (see Chapters 4 and 9 for more detailed accounts of linguistic phenomena affected by register).

One question that often accompanies studies of phraseology concerns the percentage of language that is created using semi-preconstructed phrases. Pawley and Syder (1983, p. 192) argue that "[t]he stock of lexicalized sentence stems known to the ordinary mature speaker of English amounts to hundreds of thousands." Indeed, Altenberg (1998, pp. 101–102) identified

"over 201,000 recurrent word-combinations, representing 68,000 different types of varying length and frequency"[1] in the half-million-word London-Lund Corpus of spoken English (LLC). Altenberg (1998, p. 102) suggests that "[a] rough estimation indicates that over 80 per cent of the words in the corpus form part of a recurrent word-combination in one way or another." Erman and Warren (2000, p. 50) claim that 55% of text comprises "prefabs". Ultimately, an accurate percentage is very difficult to calculate, not least because both genre and register have been shown to have a notable effect on phraseological behaviour, and so any calculation would have to take genre or register (or both) into account. Perhaps more challenging to calculations of the proportion of phraseology in English is the fact that, despite being completely incompatible with each other, "[t]he boundaries between stretches constructed on different principles will not normally be clear-cut" (Sinclair, 1991, p. 114). That is, though there is never any "blending" of stretches of text formed via the two principles, the "switch-points" are not always possible to identify. Therefore, Sinclair does not posit an estimate of the specific percentage of text comprising semi-preconstructed phrases. He is unequivocal, however, that "the idiom principle dominates" (Sinclair, 1991, p. 114).

3.3 Defining Phraseology

The idiom principle is not a methodology for revealing sets of semi-preconstructed phrases or taxonomic categories into which related phrases could be placed. It is, at its essence, a logical abstraction that describes in very general terms how language users create meaning. Sinclair (2008, p. 407), does, however note: "Phrases have never had a proper status in linguistic theory, and, as a consequence, are anomalous in descriptions." Biber et al. (2004, p. 372) agree, adding:

> despite the general consensus on the importance of multi-word units, there is surprisingly little agreement on their defining characteristics, the methodologies to identify them, or even what to call them; and, as a result, there is little agreement across studies on the specific set of multi-word units worthy of description.

A detailed survey of all of the classifications, taxonomies, approaches, and methods is neither possible in this short chapter nor necessary. However, a brief review of the approaches of two influential researchers, Alison Wray and Stefan Gries, to the problems of phraseological description will be helpful to illustrate how significant this issue of classification can be.

To begin, Wray (2002, p. 8) writes: "this large and unwieldy set of types [of phraseological item] has been carved up and categorized in innumerable ways, all of which have something useful to say, but none of which seems fully to capture the essence of the wider whole." Wray (2002, p. 9) lists at least fifty "[t]erms used to describe aspects of formulaicity." One of the main focusses of Wray's research is the argument that "much of our entirely regular input and output is not processed analytically, even though it could be." That is, although the language user's mind is likely capable — in terms of pure processing power — of adopting the slot-and-filler/open-choice model of language creation, corpus evidence appears to show that, in fact, language users do not generally employ this method of processing language.

This observation is central to Wray's (2002, p. 9) classification of "formulaic sequence" which she defines as:

> a sequence, continuous or discontinuous, of words or other elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar.

In her study of adult L1 English speakers Wray (2002) finds that formulaic language not only reduces processing load, what Sinclair (1991, p. 110) calls "economy of effort", but also aids in expressions of personal identity. These expressions of individual identity, she argues, also contribute to the maintenance of the identity of the broader community because speakers tend to share and re-use formulaic language[2].

The focus of Gries' (2008) research is quite different from Wray's, yet they both start with the same problematic necessity of defining the units they intend to study. Gries (2008, p. 10) writes, "the importance that phraseology can play in a [theoretical] framework also crucially depends, of course, on how phraseologisms are defined, which is why I devoted so much space to the question of definition." Figure 3.1 below shows the six parameters of phraseology put forth by Gries' to define "phraseologisms". These parameters, he argues, allow researchers to provide definitions of the phraseological items they investigate that are both rigorous and allow for comparisons across studies without causing undue terminological, methodological, or theoretical confusion.

Figure 3.1 Six parameters of phraseologisms, taken from Gries (2008, p. 4)

i. the *nature* of the elements involved in a phraseologism;
ii. the *number* of elements involved in a phraseologism;
iii. the *number of times* an expression must be observed before it counts as a phraseologism;
iv. the permissible *distance* between the elements involved in a phraseologism;
v. the degree of *lexical and syntactic flexibility* of the elements involved;
vi. the role that *semantic unity* and semantic *non-compositionality* / *non-predictability* play in the definition.

Gries argues: "[I]t is essential that we, who are interested in something as flexible as patterns of co-occurrence, always make our choice of parameter settings maximally explicit to facilitate both the understanding and communication of our work."

As noted, it is not possible to review all, or even most, of the approaches to and taxonomies of phraseology here. Besides, it is likely that "given the complexity of these issues, […] no single approach can provide the whole story" (Biber, Conrad and Cortes, 2004, p. 372). Cowie (1998, pp. 2–3), for example, notes in the edited collection *Phraseology: Theory, Analysis, and*

Applications, that "[t]hree major theoretical approaches are represented […], either directly, or indirectly through description or practical application." Cowie describes these three approaches as "'[c]lassical' Russian theory"; "broadly anthropological" or "cultural"; and "frequency-based". Granger and Paquot (2008) focus on two major approaches, the "Phraseological" (the foundations of which are also attributed to the Russian tradition) and the "Statistical". The former might be interpreted as generally a top-down or deductive approach, while the latter involves a bottom-up, or inductive method. These are discussed in more detail in the next section and are used here to frame the theoretical and methodological approaches to phraseological investigations following in this thesis.

3.4 Granger and Paquot: Two Approaches to Phraseology

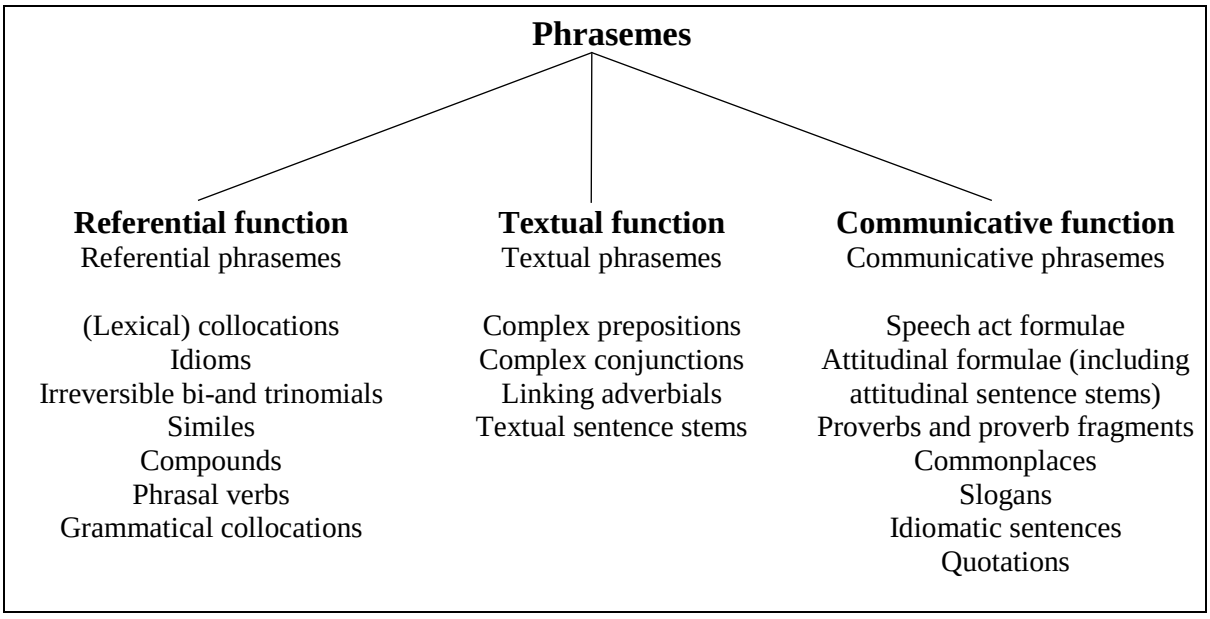
Granger and Paquot (2008, p. 27) write:

> phraseology has only recently begun to establish itself as a field in its own right. This process is being hindered by two main factors however: the highly variable and wide-ranging scope of the field on the one hand and on the other, the vast and confusing terminology associated with it.

In their effort to simplify the confusing terminology that has been associated with phraseology, Granger and Paquot (2008, pp. 28–29) begin by describing two broad approaches, "the traditional" and "the bottom-up corpus-driven", borrowing the terms "phraseological approach" and the "statistical approach" respectively from Nesselhauf (2004, as cited in Granger and Paquot 2008). Granger and Paquot's discussion continues by noting that phraseology shares "fuzzy borders" with at least four related fields, namely semantics, morphology, syntax, and discourse, and in a footnote they recognize that still other fields are perhaps equally influential, citing "phonology/prosody, sociolinguistics and psycholinguistics" (Granger and Paquot, 2008, p. 30).

In their discussion of the various categories of word combinations, and in their attempt to reconcile these two approaches to phraseology, they review in detail three phraseological typologies: Cowie's (1994, 1998) "word combinations"; Mel'čuk's (1995, 1998) "phrasemes"; and Burger's (1998, cited in Granger and Paquot 2008) "phraseological units". They also examine in detail two statistical categories: n-grams/cluster analysis; and co-occurrence analysis. Granger and Paquot conclude by outlining their own phraseological typology based on Burger's but borrowing Mel'čuk's (1995) term "phrasemes". They place eighteen phraseme types into a spectrum of three functional categories: referential, textual, or communicative. Granger and Paquot's phraseological spectrum is reproduced below in Figure 3.2.

Figure 3.2 Granger and Paquot's (2008, p. 42) Phraseological Spectrum



Granger and Paquot (2008:41) argue, however, for maintaining the distinction between distributional categories of n-gram/cluster analysis and co-occurrence analysis (i.e. the statistical approach) on the one hand, and their phrasemes, representing the linguistic approach, on the other:

> Many linguists working in the traditional framework seem to be largely unaware of the

> benefit they could derive from automatic corpus-based methods of extraction and analysis. Conversely, linguists working in the distributional framework seem not to appreciate how much they stand to benefit from the fine-grained linguistic analyses of the traditional approach.

3.5 The Frequency Approach to Phraseology

The approach taken in the current thesis is entirely corpus-based and, therefore, falls under the statistical description of approaches to phraseology. As Biber *et al.* (2004, p. 376) argue: "The usefulness of frequency data (and corpus analysis generally) is that it identifies patterns of use that otherwise often go unnoticed by researchers." However, the frequency approach is not always a simple matter of counting the number of times a phrase occurs in a corpus. This section explores three different ways frequency data can be applied to phraseological studies.

3.5.1 N-grams, lexical bundles, and chains

One of the most direct frequency-based approaches to phraseology is the extraction of n-grams from a corpus. Put simply, "N-gram analysis is a method which allows for the extraction of recurrent continuous sequences of two or more words" (Granger and Paquot, 2008, p. 38); Altenberg (1998, p. 101) is careful to add to the definition that n-grams recur "in identical form." These continuous sequences can be of any length, from two or three words (e.g. bigrams like *of course*, and trigrams like *thanks very much*) up to virtually any length, although Altenberg's data, in the form of a graph (Altenberg, 1998, p. 102), appears to have only very few 6-grams, and none that are seven or more words long. The longest recurrent word combinations directly referenced in the discussion are *but on the other hand* and *thank you very much (indeed)*.

Biber (2009, p. 276) argues that the inductive approach to N-gram analysis "makes fewer theoretical assumptions" by relying only on frequency in a corpus to identify relevant phrases.

In Biber *et al*. (2004) N-gram/Cluster analysis is used to identify categories of discourse-specific "lexical bundles", defined as "the most frequent recurring lexical sequences in a register." These bundles are subdivided into three "types": Type 1 contain verb phrase fragments; Type 2 contain dependent clause fragments; and Type three are phrasal, specifically comprising noun phrases, prepositional phrases, and expressions of comparison (Biber, Conrad and Cortes, 2004, p. 280). They show that in addition to having "identifiable discourse functions" — "(1) stance expressions, (2) discourse organizers, and (3) referential expressions" (Biber, Conrad and Cortes, 2004, p. 280) — "different registers tend to rely on different sets of lexical bundles." For example, Biber *et al*. (2004, p. 397) show that classroom teaching compared to general conversation utilizes roughly the same number of stance bundles, somewhat more discourse organizers, and substantially more referential bundles. The impact of the identification and analysis of lexical bundles lies essentially in their frequencies of use in the corpus and their functions as identified in the inductive, corpus-based, study. Biber et al. show, first, that lexical bundles are essential to communication, and, furthermore, that they are employed with varying frequencies across registers.

Somewhat similar to lexical bundles are "chains", which are defined by Stubbs (2002, p. 230) as "a linear sequence of uninterrupted word-forms, either two adjacent words, or longer strings, which occur more than once in a text or corpus." Stubbs of course acknowledges the various labels attributed to such structures including, among others, "lexical bundles" and "n-grams", but appears to apply the term "chains" to differentiate his work which is "largely methodological" and aims to identify lexical, grammatical and semantic features of the extracted phrases. In a related later work, for example, Stubbs and Barth (2003, p. 62) show that both "individual words and chains distinguish different text-types."

Stubbs' investigation into recurrent chains touches on one of the key observations made in Chapters 7 and 8 of this thesis, analysing the effects of phraseology on semantic prosody. Namely, "some of the most frequent words in the language [... are] not frequent by virtue of their single word uses [...] but because they often occur in so many set phrases or chunks" (Summers 1996, p. 262–63, cited in Stubbs 2002, p. 227). We can illustrate this point with the example of the bigram *of course*, famously discussed in Sinclair (1991, pp. 110–111), which is found 87,243 times in the BoE. These occurrences account for 56.4% of the 154,705 occurrences of the lone word *course* in the BoE. This means that whenever the single word *course* appears in a frequency list for a text or corpus, we must be aware (perhaps taking into account text-type, register, etc.) that roughly half of these occurrences are likely not referring to the path of a vehicle or river, an area of land where golf is played, a series of medical treatments, a part of a meal, etc.[3] Accordingly, Stubbs (2002, p. 230) explicitly warns that frequency lists are potentially problematic in that they can "present frequencies which are partly the result of something else: the frequency of phrases which contain the words." This issue is addressed in further detail in Chapter 8.

3.5.2 Semantic sequences and multiword units (MWUs)

Hunston (2008, p. 271) makes the important observation that "[a] sequence of words that is claimed to represent something that is 'frequent' may not itself be frequent in absolute terms […]. [I]n long strings of words, cumulative frequency may be more important than absolute frequency." Hunston (2008, 2009) reveals "semantic sequences" in corpus data by adapting a corpus interrogation technique first described by Danielsson (2007, p. 19), who calls these items multiword units (MWUs). Danielsson (2007, p. 19) describes how her methodology reveals MWUs through "a very simple sequence of actions." The first step is to identify the most

frequent lexical collocate of the node under consideration in the 4:4 span, regardless of position (as the BoE List does, for example). Then, the corpus is interrogated for all lines containing the combination of the node and this most frequent collocate (again, regardless of position) and a new list is created. The top collocate is again selected and a new concordance and collocate list is created for the three-word unit. This process repeats until no new collocates are found with a raw frequency higher than five. Danielsson illustrates the process using the word *jam* as the initial node, showing that *traffic* is the most frequent collocate in the 4:4 window. Continuing, she shows that the collocates list created from all lines containing *traffic jam* is topped by *a*, and the top collocate for *a traffic jam* is *in*. Finally, it is shown that the most frequent collocate of *in a traffic jam* is *stuck*. Danielsson has thus revealed the MWU *stuck in a traffic jam*. She notes that the next step would be to test for syntagmatic and paradigmatic variations (there are no syntagmatic variations of *stuck in a traffic jam* attested in her corpus, but she notes paradigmatic variants of *stuck*: *sitting*, *waiting*, and *caught*).

A major difference between the method described by Danielsson and that employed by Hunston is that in Danielsson's research, "'[f]unction' words are discarded" (although apparently only at the initial step; as we have seen she describes the indefinite article *a* as being the most frequent collocate of *traffic jam*, and the preposition *in* as the most frequent collocate of *a traffic jam*). Hunston (2008, p. 272), however, explicitly includes these "small words" because they are central to her analysis of textual meaning. Hunston shows that one of the primary advantages of identification and analysis of semantic sequences is that they allow for in-depth discussion of elements of co-selection that other methods might overlook. For example, while the similarities of semantic sequences and lexical bundles is undeniable, the raw frequencies of semantic sequences are most often (much) lower than frequency thresholds set in studies of

lexical bundles and would therefore be missed entirely in such studies. Semantic sequences could also be considered instances of grammatical patterning or "pattern flow", but such analyses would potentially miss both restrictions on specific elements and subtle meaning differences that emerge. For example, pattern analysis that recognizes that a verb fills a specific slot tells us nothing about restrictions on what kind of verb is required by the sequence (see Hunston (2008, pp. 278–284) for detailed discussion of the development and analysis of sequences based on the pattern 'N *that*' for further illustration).

Danielsson's procedure, by ignoring the function words (even if it is only in the first stage of the process), potentially fails to identify many salient sequences encapsulating "what is said" (Hunston, 2008) in a corpus. Danielsson's method does not identify *the worst thing that can happen*, for example, (discussed in detail in see Section 5.2.1) because *can*, though a very frequent collocate of *happen* is a member of the closed-class group of modal auxiliary verbs. In fact, the majority of N-1 collocates of HAPPEN (forty-seven of the top fifty by t-score) and its word forms are function words. For this reason, I adopted a simplified version of Danielsson's method of revealing MWUs in an early pilot study of the phraseological behaviour of the word forms of HAPPEN. Detailed discussion of this process of identifying MWUs could have been included in Chapter 5 which details the methodologies employed in this thesis, and indeed it is discussed again in Section 5.2.1. It is included here, however, because it is not only a methodology used to reveal MWUs, but it is also based on important theoretical constructs.

This technique is simplified in the sense that it includes both functional and fully lexical words at every step in the process, and thus a) does not require time-consuming manual searches and selections of only lexical words and b) does not require the human analyst to make decisions regarding the status (whether functional or lexical) of words that may fit into either category.

The method begins by calling up the frequency Picture for the node (section 5.4.2), then calling up a concordance from within this picture of the most frequent collocate at N-1. In the case of *happen*, for example, the most frequent collocate at N-1 is *to* (10,938). The frequency Picture for the concordance of *to happen* is then created. The most frequent collocate at N-1 of *to happen* is *going* (3,758). The most frequent collocate at N-1 of *going to happen* is *s* (1,338). This process continues until a phrase exhibiting "semantic unity" (Gries, 2008) is revealed or the (admittedly arbitrary) threshold of no fewer than five occurrences in the corpus is reached. Figure 3.3 below shows the first five MWUs revealed using this recursive Picture technique. Paradigmatic variations and their frequencies are shown in square brackets. The MWU *don't know what's going to happen* has a raw frequency of ninety-six; forty of these are immediately preceded by *I*. The figure shows that four collocates, *I*, we, you, just account for eight-six of the nine-six occurrences of the MWE. The remaining ten are collocates that do not meet the five-occurrence cut off (e.g. s*he*, *he*, *they*).

Figure 3.3 The first five cumulative frequency MWUs for *happen* in the BoE

[I (40) / we (20) / you (18) / just (8)] don't know what's going to happen (96)
[I (65) / we (30) / just (8) / you (7)] don't know what will happen (117)
[but (17) / and (7)] if it doesn't happen (63)
[were (3) / was (2)] worried about what would happen (12)
[it's (5)/what's (5)] the worst thing that can happen (64)

Hunston (2009, p. 145) argues, "There is a unity to the sequence, not because it is particularly frequent, […] but because it is built up of items that are statistically important to one another." For example, The MWU *worried about what would happen* in Figure 3.3 occurs only twelve times in the BoE, which is hardly enough to warrant serious investigation on grounds of frequency alone. However, in the context of uncovering its cumulative frequency, the statistical significance of each word as the MWU 'grows', it is clearly a notable sequence.

It should be mentioned, if only briefly, that the MWUs in Figure 3.3 represent only those created from the top five collocates at N-1 of *happen*, but, of course, the same technique could be applied to create a tree structure of MWUs from a single word form. Five is an arbitrary number used to illustrate the technique here, and more or fewer may be appropriate for different word forms. More importantly, a new set of MWUs could be created at each step of the process. For example, if we treat *to happen* as the node, the top five collocates at N-1 are *going* (as we have seen above), *that*, *this*, *likely*, and *it*. This (again, arbitrarily) short list reveals the MWUs, *but for **that** to happen*, *but for **this** to happen*, *of what is **likely** to happen*, and *waiting for **it** to happen*. Each of these, in turn, could then branch off, so that we could include *I don't want **that to happen***, *we can't allow **that to happen***, and so on. The full tree structure, even for a single word form like *happen*, could be enormous.

A shortcoming of this version of the method is that it is directional, unlike both Hunston's and Danielsson's techniques, and this directionality could fail to reveal potentially noteworthy MWUs. However, the method could easily be adapted to take collocates on both sides of the node into account at each stage. Software designed to automatically extract MWUs from a corpus could be programmed to find uni-directional (to the left or right) or bi-directional (identifying the most frequent collocates at either side of the node) MWUs.

3.5.3 Collocational frameworks

A defining feature of both n-grams/clusters and MWUs is that they are continuous sequences. However, discontinuous sequences allow for internal lexical or syntactic variation (cf. Sinclair 1991:111), and one of the earliest types of discontinuous sequences studied in detail are "collocational frameworks" (Renouf and Sinclair, 1991):

> Our 'frameworks' consist of a discontinuous sequence of two words, positioned at one word remove [sic] from each other; they are therefore not grammatically self-standing; their well-formedness is dependent on what intervenes. (Renouf and Sinclair, 1991, p. 129)

Specifically, the frameworks they describe comprise two high-frequency grammatical words separated by a lexical word, usually a noun or adjective. The significance of collocational frameworks lies primarily in the fact that frequency data can be applied to them in at least three different ways. First, the raw frequency of the framework itself can be informative. For example, Renouf and Sinclair show that the framework *a* + ?[4] + *of* occurs 3,830 times (tokens) in their corpus. This is very frequent compared to, for example, the MWUs discussed above. Perhaps more informative is the fact that, of these instances, 585 unique collocates (types) occupy the central slot in the corpus. The type-token ratio of these frameworks — 6:1 for *a* + ? + *of*[5] — shows that "the frameworks are highly selective of their collocates" (Renouf and Sinclair, 1991, p. 130), which, in turn, is highly suggestive of the much broader notion of the phraseological nature of meaning creation in English. Put simply, the centre positions in these frameworks are not fillable by just any noun, as in an open-choice model.

Secondly, each framework selects conspicuously different sets of collocates. Table 3.1 below shows an abridged set of Renouf and Sinclair's data for *a* + ? + *of* and *an* + ? + *of*. The table illustrates that the nouns most frequently found to complete the framework are of very different types, a point they discuss at length throughout the article.

The third application of frequency data to collocational frameworks, and perhaps the most significant, is that data like this show how important the relationships are between the patterns and the collocates completing them. For example, the data show that the triplet *a couple of* accounts for sixty-two per cent of all the occurrences of *couple* in the corpus, that *a series of*, *a*

pair of and *a lot of* each account for over 50 per cent of the total corpus instances of *series*, *pair* and *lot*, and so on (Renouf and Sinclair, 1991, p. 132).

Table 3.1 Top ten collocates and their FaC for the Collocational Frameworks a|an + ? + of, taken from Renouf and Sinclair (1991, p. 130)

a + ? + *of*		*an* + ? + *of*	
Collocate	FaC	Collocate	FaC
lot	1322	act	125
kind	864	example	77
number	762	average	73
couple	685	expression	71
matter	550	air	66
sort	451	element	58
series	438	understanding	54
piece	415	extension	45
bit	379	area	39
sense	356	hour	38

As we will see, this notion of frameworks is central to some of the phraseological analyses that follow in Chapters 7 and 8.

3.6 Canonical Form

This section discusses the argument that semi-preconstructed phrases are likely to have one form that can be considered the "prototypical" or "canonical" variant. For example, Stubbs (2007, p. 172) argues,

> as is always the case with lexical data, some patterns are so much more frequent than others that they can be taken as canonical. Sinclair proposes the strong hypothesis that for each unit of meaning there is one canonical form plus variants.

Indeed, in the interview with Sinclair conducted by Wolfgang Teubert that prefaces the 2004 publication of *The OSTI Report*, Sinclair turns to the issue of extended units of meaning and argues that "[t]heir inherent fuzziness makes them difficult objects for language teaching." The problem, he argues, is that "these new entities are fixed neither by rules nor by invariance"

(Sinclair, Jones and Daley, 2004, p. xxiv). Sinclair's proposed solution to this primarily pedagogical problem is worth quoting at length:

> The best way forward would be to construct a different model of language, a model where there would be, for each lexical item, one canonical form amid all the variation. The computer would be the tool that distilled this canonical form. One such form might be a phrase like *get in touch with*, where *in touch with* is invariable and *get* is the default collocate. There are all sorts of other verbs that could be substituted for get: *bring*, *be*, *keep*, *remain* etc.

Sinclair clarifies his position somewhat in another 2004 article "New evidence, new priorities, new attitudes." He writes, "I am using canonical form to mean the most explicit, full and unambiguons [sic] presentation of a lexical item that can be achieved." (Sinclair, 2004a, p. 298). This strong argument[6] for the canonical status of specific variations of semi-preconstructed phrases leads to a question that is central to the practical pedagogical impact and utility of the canonical form. As Sinclair suggests, "Initially, only the canonical form would be learned as a lexical item; the students would learn to recognize variants themselves" (Sinclair, Jones and Daley, 2004, p. xxiv). Though he does not say so explicitly it is clear that Sinclair intends for corpus frequency (or other frequency-based metric such as t-score) to be the primary metric in determining the canonical form: "The computer would be the tool that distilled this canonical form" (Sinclair, Jones and Daley, 2004, p. xxiv).

Sinclair's example of *in touch with* (*Ibid*.) demonstrates both the potential of recognizing a clearly dominant form as well as the importance of computer assistance in determining the canonical variant of a phrase. Table 3.2 shows the top ten N-1 collocates of *in touch with* in the BoE. The table shows that the most frequent collocate, *get*, accounts for 1,192 (20.6%) of the 5,787 instances in the corpus. The second most frequent collocate, *keep*, occurs less than half as frequently with only 533 occurrences, or 9.2% of the total. Corpus data shows that Sinclair

is clearly correct in his assessment that, based on frequency data, ***get*** *in touch with* is the canonical form of the item.

Sinclair (Sinclair, Jones and Daley, 2004, p. xxiv) argues that "there are all sorts of other verbs that could be substituted for *get*: *bring*, *be*, *keep*, *remain* etc." If we compare the collocates in Table 3.2 to Sinclair's suggestions, however, we can also now begin to see the importance of the computer's role in establishing an item's canonical form. It is often argued, for instance, that intuition is an unreliable guide to collocation and frequency (Sinclair, 1991, p. 4; Stubbs, 1996, p. 40; Hunston, 2002, p. 20).

Table 3.2 Top-ten collocates at N-1 for *in touch with* (5,787) in the BoE, showing the FaC and percent of total for each collocate

	N-1	FaC	% of total	NODE
1	get	1,192	20.6%	in touch with
2	keep	533	9.2%	
3	been	423	7.3%	
4	you	351	6.1%	
5	got	247	4.3%	
6	kept	237	4.1%	
7	be	230	4.0%	
8	stay	217	3.7%	
9	getting	165	2.9%	
10	keeping	131	2.3%	

Sinclair does not mention where these particular possible substitutions come from (i.e. his own corpus data or his intuition), but corpus evidence shows that, in fact, two of his suggestions are selected exceedingly infrequently. There are, only forty-six instances of REMAIN *in touch with* in the 450-million-word BoE, which represent only 0.8% of the 5,787 occurrences of *in touch with*. Moreover, there is only one instance (0.017%) of BRING *in touch with*:

1. or contact person **brought in touch with** the patient via telephone.

The potential pedagogical issue mentioned earlier arises not from such very infrequent instances, though. As is the case with *get in touch with*,

> [i]t happens that in most cases of varying realisations of a phrase, one of the alternatives is far more frequent than any of the others, and an obvious candidate for a canonical form, easy to teach and with the authority of a corpus behind it. (Sinclair, 2004a, p. 274).

Problems can arise, though, because focussing too much on canonical status ignores the potential salience of other forms that are worth teaching explicitly, and not left up to the student to discover on her own, as Sinclair suggests. If we use Renouf and Sinclair's collocational frameworks as an example, we can see that over-emphasis on canonical forms may overshadow other pedagogically useful information about these sequences. A convincing case could be made that the sequence *a lot of* is the canonical form of the *a* + ? *of* framework, as its 1,322 occurrences in Renouf and Sinclair's corpus are 34.5% of the 3,830 occurrences of the framework, and the second most frequent collocate — *kind* — occurs 864 times, 22.5% of the total. However, this approach omits the potentially pedagogically useful information that 53% of all occurrences of *lot* in the corpus are found in this framework. An argument could be made that *a couple of* is the canonical variant because of all the collocates that complete the framework, *couple* has the highest raw frequency to framework-frequency ratio; as we have seen, Renouf and Sinclair report that 62% of instances of *couple* in their corpus are found in the *a* + ? *of* framework. The potential significance of teaching these percentages cannot be overstated if the goal is to take a phraseological approach to lexis in the classroom.

Collocational profiles of Renouf and Sinclair's framework *too* + ? + *to* illustrate another potential problem in the classroom, namely that the canonical status of a single collocate is not always so obviously apparent in corpus data. Table 3.3 shows the top ten collocates of this framework as presented by Renouf and Sinclair (1991, p. 132) compared to the top ten in the

BoE. First, the table shows that the rankings do not agree, which alone suggests that there is no canonical form; the top two collocates, *late* and *much*, are the same in both lists but their rankings are reversed. More significantly, in Renouf and Sinclair's data, these collocates are separated by only two occurrences, and in the BoE the differ by only fifty-two occurrences, amounting to a difference of a mere 0.2 percentage points.

Even if we wanted to argue that the BoE data set is likely more accurate by virtue of the fact that it is much larger (which, in itself, would require a great deal of qualification), only the most pedantic adherence to observed frequencies and the notion of canonical form would allow us to claim that *too much to* is the canonical variant of the framework.

Table 3.3 The top-ten internal collocates for the collocational framework *too* + ? + *to* (34,268) in Renouf and Sinclair (1991, p. 132) the BoE

	Renouf & Sinclair's Data		From the BoE		
	Collocate	FaC	Collocate	FaC	% of Total
1	late	67	much	2,075	6.1%
2	much	65	late	2,023	5.9%
3	young	40	early	1,831	5.3%
4	easy	38	close	1,409	4.1%
5	small	27	young	1,176	3.4%
6	close	26	good	1,074	3.1%
7	tired	25	small	881	2.6%
8	weak	22	easy	716	2.1%
9	good	21	old	715	2.1%
10	old	18	long	572	1.7%

In section 10.3.2, an argument is made that examples like this could justify granting 'co-canonical' status. This issue is also potentially compounded when the variable collocates affect the evaluative polarity of the phrase. In section 10.3.1 an argument is made for co-canonical status when an item is observed to evaluate either positively or negatively, depending on the collocate selected. That is, when the initial functional choice to evaluate negatively is made, the negative canonical form (or variant) is selected, and when the initial choice is to evaluate positively, the positive canonical form (or variant) is selected. Although in most cases one

collocate, positive or negative, will stand out because it is considerably more frequent (as Sinclair and Stubbs both argue), relying on only one evaluative form is potentially misleading for students.

3.7 Grammatical Patterning

Discussion of grammatical patterning is included in this chapter because it represents yet another unique perspective on phraseology and the idiom principle. Hunston and Francis (2000, p. 37) write:

> The patterns of a word can be defined as all the words and structures which are regularly associated with the word and which contribute to its meaning. A pattern can be identified if a combination of words occurs relatively frequently, if it is dependent on a particular word choice, and if there is a clear meaning associated with it.

One of the reasons Hunston and Francis' grammatical patterning is regarded as such a significant addition to our understanding of meaning creation is that "[p]atterns blur the distinction between grammatical and lexical facts" (Hunston, 2002, p. 151). That is, they illustrate the co-selection of grammatical categories and lexical items in much the same way that two of Sinclair's categories of co-selection, collocation and colligation, show that groups of concrete lexical instantiations can be abstracted into grammatical classes. However, grammar patterns tend to be even more specific, and in many ways more revealing.

Corpus data show, for example, that HAPPEN is immediately followed by a preposition 48,817 times in the BoE, which is 32.7% of the 149,408 instances of HAPPEN. We can safely claim, then, that HAPPEN has a strong colligational tendency to be followed by a preposition. Pattern analysis, however, tells us that the pattern **HAPPEN *to* n**[7] occurs 16,912 times (11.3% of the total occurrences of HAPPEN) and **HAPPEN *in* n** occurs 14,342 times (9.6%). A brief look at collocates

at N+1 and N+2 in the BoE Picture for HAPPEN *to* suggests that the noun groups following *to* in this pattern are in fact realized predominantly by pronouns for specific people (*me*, *him*, *you*, *them*, *her*, *us*), and general nouns (*the people*, *the money*, *the world*, *the economy*). Some of these general world nouns are also found in the picture for HAPPEN *in* (namely *the world*, *the country*, *the west*), as well as some specific place names (*the United States*, *the Soviet Union*). But noun groups following the preposition *in* in this pattern are more frequently abstract expressions of time (*the past*, the *future*, *the next*, *the last*, *the first*); and many are in possessive structures (*my*, *our*, *your*, *their*, *his*) of abstract time nouns (*past*, *life*, *future*). The observation that the noun groups following the prepositions are subtly different is possible primarily because of a more holistic patterning approach. This observation might be missed in a more general analysis of the lexical item that distinguishes only collocates that are nouns and those that are prepositions.

There are three ways in which patterns are associated with meaning. Hunston and Francis (2000, p. 3) write, "firstly […] in many cases different senses of words are distinguished by their typical occurrence in different patterns." For example, Hunston and Francis (1998) show that nine patterns of the verb CONSIDER are associated with its three main senses in the CCED. The first sense, meaning roughly to hold an opinion about someone or something, can be realized by the patterns **V n to-inf**, **V n n/adj**, **V that**, and **V n as n/adj**. Lines 2 to 5 below are four of Hunston and Francis' (1998, p. 47) examples of these four patterns:

```
2. it clearly unconstitutional and we consider it to be a flagrant violatio
3.   indecently flattered He does not consider himself a celeb, retains the
4.      the social costs If we do not consider that the costs are worth pay
5.           why the region should be considered as a special case foe extr
```

The second association between pattern and meaning can be observed in the fact that "words which share a given pattern tend also to share an aspect of meaning." For example, Hunston

and Francis (1998, p. 52) list forty-eight of the most frequent verbs associated with the pattern **V *as* n**, which they then arrange into six meaning groups. For example, the first meaning group for verbs fitting this pattern are those "concerned with working at a job" such as *work*, *qualify*, *train*, *freelance*, *moonlight*, etc. They provide two corpus examples to illustrate, which are reproduced (abridged) below:

6. I'm practising as a doctor
7. he is expected to resign as governor

They note that "[i]n all the patterns we have identified (Francis *et al* 1996), a similar division into meaning groups can be made."

Hunston and Francis (1998, p. 69) explain the third way in which pattern and meaning are potentially associated is that

> the patterns themselves can be said to have meanings, and there is some evidence that the use of a lexical item with a pattern that it does not commonly have is a resource for language creativity and, possibly, for language change.

They provide the example of the pattern **V n *as* n** which frequently uses verbs like *consider*, *describe*, *interpret*, etc., which are all used in expressions of opinion as opposed to fact. The following examples from the BoE illustrate:

8. MP's said they still **considered** America as Islam's enemy number
9. against her. She **described** Palmar as `evil" and `weirder than
10. his unfortunate tendency to **interpret** dissent as treachery, his biggest

Hunston and Francis (2000, pp. 106–107) then cite an example of a writer who uses the verb *discover* in this pattern:

11. Society must be predisposed to panic about crimes. There has already to be a tendency to **discover** crime as the cause behind the worrisome social ills.

The meaning of *discover* (to uncover a fact) appears to be at odds with the meanings of the verbs most frequently found in this pattern (to express opinions). Therefore, they paraphrase the sentence as, "'People think it is a fact that crime is the cause of social ills, but actually it is only an opinion'" (Hunston and Francis, 2000, p. 107).

It is worth noting that Hunston and Francis are not alone in their goal of showing how "[p]atterns and lexis are mutually dependent" (Hunston and Francis, 2000, p. 3). For example, Goldberg (2003, p. 219) defines "constructions" as "stored pairings of form and function, including morphemes, words, idioms, partially lexically filled and fully general linguistic patterns." A principle difference between constructions and patterns is that "[a]ny linguistic pattern is recognized as a construction as long as some aspect of its form or function is not strictly predictable from its component parts or from other constructions recognized to exist" (Goldberg, 2003, p. 219). This notion of non-predictability, however, is not a factor in any of Hunston and Francis' discussions of pattern grammar.

Another key difference between the two theories is that patterns "are observable from investigation of an electronically-stored corpus of written and spoken texts" (Hunston and Francis, 2000, p. 1). Constructions, on the other hand, do not require corpus investigation to be revealed. In theory, a single occurrence can be identified as a construction. Of course, the analyses of CAUSE reported on in Chapter 9 could have been undertaken from the perspective of construction grammar. Stefanowitsch and Gries (2003) perform "collostructional analysis" on the three constructions they observe CAUSE to occur in. They find that "[c]learly, *cause* has a 'negative prosody' in all three constructions; however, there are fundamental differences between the three constructions with respect to the exact type of negative result" (Stefanowitsch and Gries, 2003, p. 222; see also Gries and Stefanowitsch, 2004).

A final important distinction between pattern grammar and construction grammar relevant to the current study is that pattern grammar "does not […] take a theoretical stance on the mental processing of grammar" (Hunston, 2011, p. 123), while construction grammar is postulated as a direct "contrast to the mainstream 'generative' approach to language" (Goldberg, 2003, p. 219). Since the current study is corpus-based and, as such, is not fundamentally concerned with psycholinguistic claims, investigations from the perspective of pattern grammar seemed fitting.

Furthermore, even though every aspect of co-selection from fixed phrases to collocations and long phraseological items can be described in terms of patterning (Hunston and Francis, 1998, p. 63), in this thesis, the semi-preconstructed phrases *the* ADJ *thing that can happen*, and *a|an* NOUN *waiting to happen*, analysed in Chapter 8, are not treated as grammatical patterns. This is primarily because, as Hunston and Francis (2000, p. 37) argue, "complementation patterns are usually the most interesting facts about verbs." The complementation patterns associated with HAPPEN, though, do not appear to have a strong effect on its semantic prosody. Indeed, the pattern 'HAPPEN to-inf' is explicitly removed from concordances analysed by Bublitz (1996) because it expresses the "by chance meaning" of the verb, which does not, he argues, have a negative semantic prosody. Further, the phrases analysed here were extracted from corpus data using variations of the MWU methodology (see Section 5.2.1) adopted in early pilot studies of the phraseological behaviour of HAPPEN. The techniques employed in these studies focussed almost exclusively on collocates to the left of HAPPEN in building up phrases based on cumulative frequency. Finally, even when collocates of the single-word HAPPEN are examined for evidence of semantic prosody, we are most often concerned with *what* happens, i.e. the grammatical subjects of the verb, which are found primarily to its left.

On the other hand, complementation patterns associated with CAUSE were analysed as part of

the current research (see Chapter 9). The pattern grammar approach seems appropriate in this case because studies focussing on the semantic prosody of CAUSE tend to focus on that which is caused, i.e. the grammatical object(s) of the verb (although both subjects and objects tend to be negative). The primary reason that the grammatical patterning of CAUSE is scrutinized here, however, is because Susan Hunston (2007) provides twelve neutral instances of CAUSE, from a register-specific sub-corpus of the BoE comprising issues of The New Scientist Magazine. However, only two of CAUSE's grammatical patterns are represented in this short concordance. Hunston chose these twelve lines not to illustrate the effects of register or patterning, but simply to show that even items exhibiting a strong semantic prosody, as CAUSE does, do not always evaluate according to that prosody. Chapter 9 explores whether neutralizing the negative semantic prosody of CAUSE in Hunston's sample is an effect of register, patterning, both, or neither.

3.8 Conclusion

This chapter has discussed the theoretical foundation that informs the analysis of the effects of phraseology and grammatical patterning on semantic prosody that follows in Chapters 7, 8, and 9. For the purposes of this thesis, analysis of the phraseological behaviour of HAPPEN and the grammatical patterning of CAUSE seemed the most suitable approaches to uncovering potential problems with claims that these items have negative semantic prosodies.

For example, *the worst thing that can happen* is an MWU identified using a variation of methods employed by Danielsson (2007) and Hunston (2008). It does not occur in the corpus very frequently, but it is salient not least because of the cumulative frequency of the items comprising the phrase. Further, *worst* as a collocate of HAPPEN appears in the Picture apparently

only by virtue of being a part of this phrase (cf. Stubbs' (2002) discussion of lexical "chains"). Similarly, the prevalence of *thing* and *things* was noticed during early studies of frequency-based collocational profiles of HAPPEN. Further examination of these profiles led to uncovering the frequent collocates of *things happen*, most notably *these* and MAKE. The phrase *a|an accident/disaster waiting to happen* was selected for study in this thesis because in an early pilot study of MWUs *waiting* was found to be the sixth most frequent collocate of *to happen* in the BoE, and *accident* and *disaster* are the first and second ranked collocates of *waiting to happen*. In contrast, there seem to be far fewer salient clusters, discontinuous frameworks, and MWUs associated with CAUSE. However, the patterning of CAUSE is easily observable in corpus data.

In the chapters that follow, these structures will be referred to simply as "phrases" and "patterns". The previous research presented in this chapter and its theoretical foundations all inform that which follows, though no attempt is made to ally the current research with any previous phraseological taxonomies or methods beyond, of course, the general frequency approach.

Notes

[1] Admittedly, many of the items identified are "mere repetitions or fragments of larger structures (e.g. *the the, and the, in a, out of the*) (Altenberg, 1998, p. 102). When Altenberg restricts the data to three-word combinations and eliminates unintentional repetitions, he is left with 6,692 tokens representing 470 types for analysis.

[2] Wray continues in her investigation into formulaicity by examining the role of formulaic sequences in second language acquisition and language loss, focusing on the re-acquisition of language by patients suffering aphasia.

[3] These meanings have been paraphrased from the Collins Dictionary Online which has at least

twenty-eight meanings and usages in the British English section alone for *course*. The first meaning in the learner English section of the page, however, is: "often used in the expression 'of course'."

[4] The question mark in the notation represents the variable lexical word. In the case *a* ? *of* we see *a* ***bit*** *of*, *a* ***lot*** *of*, *a* ***couple*** *of*, etc. in corpus data.

[5] Other frameworks have much smaller type:token ratios of only two or three to one, indicating a very high degree of restriction.

[6] In another discussion of the importance of canonical form, however, Sinclair (2004a, p. 281)hedges his claims somewhat: "a canonical form of the lexical item can be proposed, which will […] be **one of the commonest forms** of the lexical item" (emphasis added).

[7] Grammatical patterning notation follows Hunston and Francis (1998, 2000). The pattern **CAUSE n**, for example, can be paraphrased as any of the word forms of CAUSE immediately followed by a noun group. "Where a preposition, adverb, or other lexical item is part of a pattern, it is given in italics to indicate that it is a lexical item rather than a code" (Hunston and Francis, 2000, p. 45); for example, Section 9.3.3 examines **CAUSE *by* n**, represents any of the word forms of CAUSE followed by the word form *by* followed by a noun group. An element of the pattern is capitalized if it is the focus of the current discussion/investigation; '+' is not used between pattern elements. For a full list of notations an explanations, please refer to Hunston and Francis (2000, pp. 44–48).

CHAPTER 4:
GENRE AND REGISTER

4.1 Introduction

This chapter looks briefly at two closely related but discrete perspectives of analysis of text varieties, "register" and "genre", in order to provide a theoretical foundation for the corpus studies discussed in Chapter 9. The chapter begins in Section 4.2 by attempting to define each term and distinguish between them, because they are often used by corpus researchers without comment or definition. This section looks at some examples of how well-known researchers (Biber and Conrad 2009; Halliday 2014; Swales 1990) have used the terms register and genre and ends with a concise catalogue of features of both areas of investigation as determined by Biber and Conrad in their book *Register, Genre, and Style* (2009). Section 4.3 discusses some specific linguistic factors, both lexical and grammatical, that have been shown to be affected by specific registers and genres. Finally, Section 4.4 looks at contrasting claims in the literature regarding the potential effects of register and genre on semantic prosody. The chapter concludes with a brief discussion of the approach adopted in the research presented in Chapter 9.

4.2 Defining Register and Genre

This section attempts to define the terms "register" and "genre" in order to distinguish between the two and to begin to orient the results of research presented in Chapter 9. Also mentioned briefly in this section are the terms "text-type", "domain", and "style". These terms appear here largely because the researchers cited often include them in their discussions of register and genre. However, it was decided that in-depth discussions of text-types, domains, and styles would unduly complicate this discussion and take emphasis away from the more widely used notions of register and genre.

Swales argues, "it is only comparatively recently that genre has become disentangled from register" (Swales, 1990, p. 40). Biber and Conrad (2009:21), however, warn that this disentanglement is not yet complete: "It is important to be aware that there is no general consensus concerning the use of *register* and related terms such as *genre* […]" (original emphasis). For example, Stubbs (1995) claims that the negative semantic prosody of CAUSE persists across all genres represented in the LOB[1]. However, the manual for the LOB, originally published in 1978 and now held online[2], contains only one instance of the term "genre" (and none of "register"). Instead, there are more than 160 instances of the terms "(sub)category/(sub)categories". Therefore, it appears that Stubbs is retroactively applying the term "genre" (though apparently not indiscriminately because he provides the examples "newspapers, reports, academic articles, novels") to what the creators of the LOB corpus chose to describe only as categories of text.

Hoey's *Lexical Priming: A New Theory of Words and Language* contains thirty-four instances of "genre/s" (one of which is found in the index, and two in the bibliography), but "register" occurs only once (as part of an article title in the bibliography). Additionally, both the hyphenated "text-type" and unhyphenated "text type" are found only once each, first when he characterizes his corpus as "text-type-specific" (Hoey, 2005, p. 72), and again when he writes, "I have stressed repeatedly throughout this book that primings are not in principle generalised across all text types, genres and domains" (Hoey, 2005, p. 174). This last term, "domain", occurs 34 times in the book (once in the index), most often in very close proximity to genre; for example, "priming is posited to be genre and domain specific" (Hoey, 2005, p. 111); and "every word is characteristically primed for a range of genre, domain and situationally-specific features" (Hoey, 2005, p. 165). Hoey does not, however, offer definitions of any of these terms,

nor does he suggest how or why he differentiates them.

Louw and Chateau (2010:757) attribute to Partington (1998) the claim that "newspaper English" is a *genre* which contains most other genres. Partington does discuss the notion of genre as it relates to "general English" in his justification for the design of his newspaper corpus but quickly switches from "genre" to "text type" (though, again, no formal definitions are offered): "They [newspapers] also have a few particular advantages over other text types. […] Moreover, it must be remembered that newspapers consist of not one but a large number of different text types" (Partington, 1998, p. 13).

Most attempts to explicitly define genre centre on the language user's communicative purpose. For example, Dudley-Evans (1994, p. 219) writes:

> The term **genre** was first used in an ESP context by Tarone et al. (1981) in an article that investigated the use of the active and passive forms in astrophysics journal articles. That article established the principle that within the conventions of the genre studied it was the writer's **communicative purpose** that governs choice at the grammatical and lexical levels [original emphasis].

Swales (1990:33), begins by acknowledging that genre is "slippery" and "a fuzzy concept" and devotes a section of a chapter to a very detailed explication of the term. The section, "A working definition of genre" (Swales, 1990, pp. 45–58) ends with a paragraph-length summary that is worth quoting in its entirety here, not least because it, too, focuses on communicative purpose as a defining aspect of genre:

> A genre comprises a class of communicative events, the members of which share some set of **communicative purposes**. These purposes are recognized by the expert members of the parent discourse community and thereby constitute the rationale for the genre. This rationale shapes the schematic structure of the discourse and influences and constrains choice of content and style. **Communicative purpose** is both a privileged criterion and one that operates to keep the scope of a genre as here conceived narrowly focused on comparable rhetorical action. In addition to purpose, exemplars of a genre

> exhibit various patterns of similarity in terms of structure, style, content and intended audience. If all high probability expectations are realized, the exemplar will be viewed as prototypical by the parent discourse community (Swales, 1990, p.58, empasis added).

Bhatia (1997) and Flowerdew & Dudley-Evans (2002) also define genre in terms of communicative purpose. Bhatia (1997, p. 191), however, argues against "the notion of pure genres," suggesting instead that genre-mixing is an under-recognized phenomenon. Similarly, while a formal structure of moves and steps defines many, if not most, genres, Flowerdew and Dudley-Evans (2002, p. 465) argue "that a move approach is less valid for [some] academic genres […] where there is great variety in the moves adopted by different writers."

Register is defined quite differently. For Couture (1986, cited in Swales, 1990, p.41), "genres (research report, explanation, business report) are completable structured texts, while registers (language of scientific reporting, language of newspaper reporting, bureaucratic language) represent more generalizable stylistic choices." Here we begin to see that genres tend to be differentiated at a textual level, whereas registers seem to be identifiable by more abstract lexico-grammatical, meaning-creating structures. It could be suggested that genre creates a framework and register provides appropriate language to fill in that framework.

Halliday argues, in the 1975 article "Language as Social Semiotic", cited here as it appears in *The Discourse Study Reader*: *Main currents in theory and analysis* (Angermuller, Maingueneau and Wodak, 2014), that "register is the semantic variety of which a text may be regarded as an instance" (2014, p. 266). More specifically, Halliday (2014, p. 267) contends that "[a] register can be defined as the configuration of semantic resources that a member of the culture typically associates with a situation type." Halliday (2014, p. 266) also describes the situation type as the "social context" and is defined by the interaction of three dimensions, field, tenor, and mode,

that allow us to describe how language creates meaning as social activity:

> The field is the social action in which the text is embedded; it includes the subject-matter, as one special manifestation. The tenor is the set of role relationships among the relevant participants; it includes levels of formality as one particular instance. The mode is the channel or wavelength selected, which is essentially the function that is assigned to language in the total structure of the situation; it includes the medium (spoken or written), which is explained as a functional variable.

Halliday stresses that registers can be identified by surface linguistic factors (i.e. lexical and grammatical conventions employed), but the register itself is not a product of these linguistic factors. Instead, registers are reflections of the "selection of meanings that constitutes the variety to which a text belongs" (Halliday, 2014, p. 267). Register is an abstract "conceptual framework" that describes how meaning is created in different situations.

Biber and Conrad (2009:15) are at pains to emphasize that they "regard genre, register, and style as different approaches or perspectives for analyzing text varieties, *not* as different kinds of texts or different varieties. In fact, the same texts can be analyzed from register, genre, and style perspectives" (Biber and Conrad 2009:15; original emphasis). They stress that both register and genre analysis "[include] description of the purposes and situational context of the text variety" but genre analysis is distinct in that it "[focuses] on the conventional structures used to construct a complete text within the variety" (2009:2).

According to Biber and Conrad (2009, p. 16) rhetorical organization is one of the defining characteristics of genre, as shown in Table 4.1 below. These conventional structures include another frequently cited element of genre, the "rhetorical moves" and "steps" mentioned briefly above. Genre analysis, then, often involves a "moves analysis" that judges the degree to which the exemplar text negotiates the expected rhetorical organization of the genre.

Table 4.1 Defining Characteristics of register and genre, taken from Biber and Conrad (2009, p. 16)

Defining characteristic	Register	Genre
Textual Focus	sample of text excerpts	complete texts
Linguistic characteristics	any lexico-grammatical feature	specialized expressions, rhetorical organization, formatting
Distribution of linguistic characteristics	frequent and pervasive in texts from the variety	usually once-occurring in the text, in a particular place in the text
Interpretation	features serve important communicative functions	features are conventionally associated with the genre: the expected format, but often not functional

For example, Swales (1990) discusses the "Create a Research Space" (CARS) model of research article (RA) introduction sections. The CARS model comprises three moves, each of which is realized by one or more steps. Figure 4.1 below, taken from Swales (1990, p. 141) illustrates the basic move and step structure of the CARS model.

Figure 4.1 "A CARS model for article introductions", from Swales (1990, p. 141)

Move 1	Establishing a territory
Step 1	Claiming Centrality
	and/or
Step 2	Making topic generalization(s)
	and/or
Step 3	Reviewing items of previous research
Move 2	Establishing a niche
Step 1A	Counter-claiming
	or
Step 1B	Indicating a gap
	or
Step 1C	Question-raising
	or
Step 1D	Continuing a tradition
Move 3	Occupying the niche
Step 1A	Outlining purposes
	or
Step 1B	Announcing present research
Step 2	Announcing principle findings
Step 3	Indicating RA structure

Swales' detailed explanations of each of the moves and steps, including examples of each, includes an extended discussion of linguistic factors that characterize them (Swales, 1990, pp. 149–166). For example, he begins by outlining types of citations found in Move 3 ("integral" or "non-integral"), including reporting verb tense and aspect usage. Swales continues with an extended discussion of the use of negative quantifiers (*no, none, little,* etc.), other methods of employing lexical negation (*fail, lack,* etc.), and so on[3] in Move 2.

A full summary of Swales' analysis is not especially relevant to forming the foundations of the research presented in this thesis. In any case, it could be argued that, according to Biber and Conrad's Defining Characteristics of Register and Genre (Table 4.1 above), the linguistic factors Swales analyses in research article (RA) introductions are, in fact, register-specific and not genre-specific.

The approach to the analysis of corpus data presented in Chapter 9, then, is generally register specific in the sense that the studies concern the language of newspapers, academic writing, scientific writing, and so on (Couture 1986, cited in Swales, 1990, p. 41) and do not take into consideration the rhetorical organisation of texts comprising the corpora examined (Biber and Conrad, 2009, p. 16). Further, as Table 4.1 above has shown, register characteristics tend to be functional as opposed to conventional, and it has been argued (see Section 2.4.5) that semantic prosodies represent a speaker/writer's initial functional choice when selecting a lexical item.

Analyses presented in Chapter 9 rely entirely on the organization of the twenty sub-corpora, shown below in Table 4.2, that together comprise the *ca.* 450-million-word Bank of English, which is used principally in this thesis. Although it does not appear that either genre or register were used as guiding principles in the design of the BoE (see Section 5.3.1 below for a more

detailed discussion on the specific design criteria and contents of the BoE), the sub-corpora comprising the BoE do contain what can, in many cases, be considered register specific language. The language of the various news sub-corpora, for example, can be contrasted against the language of academic books, spoken English, and so on. Although future studies would likely benefit from a more granular approach — for example, by distinguishing the language of 'hard' news from sports, business, and entertainment reports that also appear in newspapers — it was decided that the BoE in its current configuration offers enough samples of diverse language use to test the effects of register on semantic prosody.

Table 4.2 Sub-corpora of the *ca.* 450-million-word Bank of English, as presented when accessed by Telnet

Sub-Corpus	Description	Number of words
usacad	US Academic Books	6,341,888
usephem	US Ephemera	3,506,272
newsci	The New Scientist	7,894,959
npr	National Public Radio	22,232,422
sunnow	The Sun and News of the World	44,756,902
brbooks	British Books	43,367,592
brmags	British Magazines	44,150,323
guard	The Guardian	32,274,484
econ	The Economist Magazine	15,716,140
bbc	British Broadcasting Company	18,604,882
usspok	US Spoken	2,023,482
wbe	Business English	9,648,371
strathy	Corpus of Canadian English	15,920,137
oznews	Australian News	34,940,271
brephem	British Ephemera	4,640,529
usbooks	United States Books	32,437,160
usnews	United States News	10,002,620
indy	The Independent	28,075,280
times	The Times/Sunday Times	51,884,209
brspok	British Spoken Englsih	20,078,901
	Total number of words	448,496,824

Table 4.2 above shows the sub-corpus names as they appear in the Telnet window and in the

same order in which they are listed, a short description of the corpus, and the number of words. The BoE software allows the user to select both individual sub-corpora or to combine any number of them as the study requires. Table 4.3 shows the general registers examined, the sub-corpora used to represent those registers, the number of tokens in each sub-corpus expressed in millions of words, and the total number of tokens in each register examined. Throughout the analyses presented in this thesis, "BoE" continues to refer to the full 450-million-word Bank of English corpus.

It should be noted that **newsci**, the main focus of comparison in Chapter 9, is a collection of full issues of New Scientist and therefore comprises many distinct genres, subgenres, registers, and sub-registers of scientific journalism and journalistic forms—including science news, reports on technological advances, emerging medical treatments, lifestyle choices, events (science-based lectures and classes), job listings, letters to the editor, etc.

Table 4.3 Combined BoE sub-corpora comprising general registers examined in this thesis; tokens expressed in millions of words

News	Tokens
sunnow	44.7
guard	32.2
econ	15.7
oznews	34.9
usnews	10.0
indy	28.0
times	51.8
Total	217.6

Books	Tokens
brbooks	43.3
usbooks	32.4
	75.8

Newsci	Tokens
newsci	7.8
	7.8

Spoken	Tokens
brspok	2.0
usspok	20.0
	22.1

Usacad	Tokens
usacad	6.3
	6.3

However, the study presented in Chapter 9 is directly inspired by, and directedly references, Hunston's (2007) small **newsci** concordance. As noted above, an interesting idiosyncrasy of this concordance is that the twelve lines comprising it use only two of the grammatical patterns of CAUSE. In order to directly address the potential effects these patterns have on semantic prosody, it was decided to focus on the same sub-corpus for comparison to the general BoE.

4.3 Linguistic Factors Affected by Register and Genre

Before turning to a specific discussion of the potential effects of register and genre on semantic prosody, this section discusses other linguistic factors that have been demonstrated to differ by register and genre. The phrase 'linguistic factors' is being used very generally here and includes a broad range of grammatical/syntactic and lexical phenomena. It is used in opposition to schematic/structural/organizational factors which include the moves, steps, stages and so on discussed above.

To begin, Biber and Conrad (2009, pp. 53–54) describe characteristics that allow for linguistic analysis of both register and genre[4]. Specifically, they differentiate among "register features", "register markers", and "genre markers". Register features are lexical or grammatical structures that are both pervasive and more frequent in the target register, though they may be found in other registers to a lesser degree. They note, for example, that the passive voice is a register feature of academic writing; this means simply that even though verbs in the passive voice are employed in virtually all registers, they are notably frequent in academic writing. In contrast, "register markers" are linguistic constructions that do not normally occur in other registers at all. For example, Biber and Conrad (2009, p. 53) note that American sports fans will immediately know the difference between the phrases "it's three and two" and "it's third and four." The former refers to baseball, the latter to American football, and they are never confused by experienced English speakers familiar with these sports. Genre markers, however, are (generally) formulaic expressions that tend to be "conventional rather than functional" (Biber and Conrad, 2009, p. 53) and mark structural divisions in a text. As such, genre markers usually occur only once in that text. They give the example, among others, of a business letter beginning with *Dear Sir*, or a prayer ending with *Amen*. Both are crucial elements of their respective

genres, but neither are particularly frequent.

Researchers have, however, studied relationships between linguistic factors and the structural elements of specific genres. For example, Paltridge (1994) examines whether linguistic criteria define moves within a genre, and concludes that "structural divisions in texts" are demarcated by "*cognitive* boundaries in terms of *convention*, *appropriacy*, and *content*" (original emphasis) and that move transitions are in fact not linguistically defined. Swales and Najjar (1987, p. 184) focus primarily on the move structure of RA introductions but do comment on "…the surface realization of the switch to first person (a typical marker for the onset of Move 4)." They also briefly discuss tense shifts and the use of passive verbs in RA introductions. These findings could be argued to more generally represent register effects, not genre-specific language. Their principal goals, however, are not to uncover the linguistic factors central to the sub-genre investigated. Instead, they aim to demonstrate the ways in which published RAs do not generally agree with prescriptive style guides in providing explicit statements "announcing principle findings" (APF). Further, they show that introductions differ across disciplines in the choices surrounding the inclusion or omission of both Move 4 and the APF, and finally that the trends they examine have changed over time.

Flowerdew and Dudley-Evans (2002:465) claim that "[i]t is no longer valid to present a study that focuses on the moves that a writer uses without consideration of the role of the writer in the discourse community and the expectations of that community." As such, in addition to the move structure of editorial letters sent to potential contributors to an international academic journal, they analyse "the linguistic features and, in particular, how they realize the interpersonal dimension of the communication" (Flowerdew and Dudley-Evans, 2002, p. 469). Specifically, they focus on items used in "maintaining good relations and avoiding face-

threatening communicative behaviour" (Flowerdew and Dudley-Evans, 2002, p. 483) such as personal pronouns, *think*, and politeness strategies (i.e. modal auxiliaries and items like *sorry*, *afraid*, etc.). They find that, ironically, while these strategies are employed as face-saving techniques directed at the addressees of the letters, they often prove to be a source of confusion for the reader.

There are many studies, however, that focus on purely linguistic effects of register or genre. In a very early comparison of "background articles in the daily press" in English, French, and Hebrew, Weizman (1984:40) found that "quotation marks combine with other function markers in discourse units of various sizes to imply certain nuances of the reporter's attitude." Her findings have implications for both translators, who, she argues, should be aware of the degree of implicitness expected by readers in the target language, and teachers who hope to anticipate register or genre transfer errors in student language output.

Malcolm (1987:38) identifies writers' "obligatory constraints [… and …] strategic choices" of verb tense in scientific articles. The ramifications are primarily pedagogical; "a list of 'uses' to memorize […] is much easier to forget than a general understanding of how temporal references affect tense choices" (Malcolm 1987:41). In addition, Malcolm suggests that students would benefit from the knowledge that tense choices are often made to further rhetorical goals. Similarly, Thompson and Yiyun (1991) present a preliminary classification of over 400 reporting verbs used in academic writing, focussing on how citations express evaluation. Again, the goal is largely pedagogical, beginning with "the simple aim of identifying a specific subset of the lexical items which we felt it would be useful for our students to know: namely, the verbs used in citations" (Thompson and Yiyun, 1991, p. 365).

Hedging is an important facet of many genres and registers, and Hyland (1996) confirms that they are extremely frequent in scientific writing: “Hedges represent more than one word in every 50 in my corpus” Hyland (1996, p. 259). Hyland investigates the contexts and motives, both social and linguistic, in which various hedges are employed and shows that they vary widely. Yet again, the conclusions are pedagogically focussed. Hyland (1996, p. 278) argues: “The RA is the key genre in academic disciplines, and familiarity with its conventions, including the ability to recognize and use hedges accurately, is vital” both for beginner L1 writers and, perhaps more importantly for L2 learners.

Swales et al. (1998, p. 117) takes “a preliminary look at the forms and functions of imperatives in scholarly articles across a range of fields” and concludes that imperative forms tend to be found within sections containing the primary argument of the paper, and that they are “complex textual signals by which academic writers manipulate various rhetorical strategies” (Swales *et al.*, 1998, p. 99). Once more, the primary focus is pedagogical as most academic style guides provide little or no guidance to novice and non-native English-speaking scientific researchers and writers.

Lindemann and Mauranen (2001) look at patterning that includes the word form *just* in the general register of academic speech (using the MICASE[5] corpus). They found that “the most common use of *just* in this data is as a hedge, or minimizer, which can reasonably be seen as a type of hedge” (Lindemann and Mauranen, 2001, p. 464). They also highlight the importance of the phonetic form, “specifically, a reduced form of *just* appears to be particularly appropriate for a mitigating function” (Lindemann and Mauranen, 2001, p. 473), as opposed to a stressed form, which may indicate impatience on the part of the speaker. They provide the example, “if you’d just wait a minute …”, and argue that stressing *just* might seem impolite (again, of

particular importance to the L2 learner).

Thompson (2009) looks at both the move structures of PhD thesis literature reviews and frequent nouns used in this subgenre. Specifically, Thompson examines the patternings surrounding the nouns *evidence*, *problem*, and *model*, "in order to establish what is distinctive about the literature review as a (sub)genre and to identify some of the strategies that are used in these reviews" (Thompson, 2009, p. 50). The patterns reveal that "the writer's voice is the dominant voice in the text" (Thompson 2009:65), despite the expected averral to expert authors and researchers in the literature review. In addition, it is demonstrated that these patterns vary across disciplines.

Groom (2009, p. 125) shows that "phraseology and epistemology are indissolubly interlinked." Specifically, Groom suggests that "reiterativeness" phraseologies associated with "closed-class" words are especially informative in distinguishing two disciplines, History and Literary Criticism. He shows, for example, that patterns associated with the prepositions *of*, *against*, *beyond*, and *upon* differ in frequency between the two disciplines examined and suggests "a working hypothesis to pursue in further studies" of other disciplines.

Hyon (2011) studies evaluation in the occluded genre of university faculty tenure and promotion letters. Hyon takes a lexical approach, looking in general at frequencies of negative and positive words and phrases. She writes: "[o]f particular interest are letter writers' linguistic choices for evaluating faculty in teaching, research and service (a common tripartite frame of faculty performance) and what these choices suggest about how faculty work is idealized and how negative assessments are negotiated" (Hyon, 2011, p. 392). She finds that negative evaluations of faculty members are most often mitigated, suggesting that face-saving and

maintaining positive working relationships between the letter writer and the subject of the evaluation are important considerations.

In conclusion, these studies tend to focus on the linguistic factors central to the social/interactive/communicative natures of the genres or registers in question. Indeed, in most cases, the very reason the studies are register-/genre-specific is to highlight the interpersonal facet of the discourse in question. Additionally, virtually all of the studies cited here (and many more besides), convey explicitly pedagogical goals or implications. In sum, it is frequently argued that students — especially L2 learners — be made explicitly aware of linguistic factors that are affected by register or genre; 'mis-use' of linguistic factors in frequency or form may lead to confusion, misunderstanding, or even complete failure to achieve the desired communicative purpose or to create the desired meaning.

4.4 Semantic Prosody and the Effects of Register and Genre: Contrasting Claims

This section turns to researchers' claims regarding the effects of register on semantic prosody. The question of how register or genre might affect semantic prosody is not as straightforward as it may seem. Sinclair makes no claim regarding potential effects of genre or register on the semantic prosody of a lexical item. Sinclair (1991, p. 110) does argue, though, that register is a limiting factor directly affecting co-selection via the open-choice principle: "[o]nce a register choice is made, and these are normally social choices, then all slot-by-slot choices are massively reduced in scope or even, in some cases, pre-empted." He does not, however, comment on how register might affect co-selection via the principle of idiom. Nor do the articles outlining the categories of co-selection comprising the lexical item (see Section 2.2) contain any reference to register or genre.

Louw (1993, p. 159), however, makes the very strong, though purely theoretical (as opposed to empirical/corpus-based), claim that semantic prosody is not genre-/register-specific; he contends that "'contagion[6]' is a general linguistic phenomenon which pervades every type of language. In other words, no amount of genre-based or register-based study in this century could ever have revealed its presence". In a much later paper, however, Louw and Chateau (2010) argue that the negative semantic prosodies of the related lexical items CAUSE, BRING *about*, and GIVE *rise to* are "smoothed" (meaning they collocate primarily with neutral, rather than negative, items) in a corpus of writing from the 'hard' sciences. One reason suggested is that "[t]he verb 'cause' is difficult to replace: synonyms are often not single words but multi-word expressions […] and, in scientific writing, a single-word will often be preferred to a phrasal verb or multi-word expression" (Louw and Chateau, 2010, p. 763). They also suggest that the sub-genre of the methods and materials sections of scientific research articles tend to contain more instances of incomplete contexts of situation, which in turn is claimed to shift the prosody of CAUSE from negative to neutral.

Partington (1998) at first seems to agree with Louw (1993) that semantic prosody is a feature of general language, adding, however, that some registers may display semantic prosodies more readily. Both Partington (2004, p. 134) and Stubbs (1995, p. 5) argue that semantic prosodies (especially negative ones) are more frequent, for example, in newspaper reportage. Partington (1998, p. 77) suggests, therefore, that "[i]t may […] be worthwhile to begin looking at texts which are likely to exploit prosodic effects — newspapers, political language, advertising, etc. — to isolate potential candidates and subsequently follow this up by examining corpus data." However, this approach is essentially the opposite of Stubbs', who, as mentioned above, notes first the general negativity of the collocates of CAUSE, and then confirms it across registers.

Stubbs (1995) makes no explicit claims about whether semantic prosody itself is register- or genre-specific, but he does note in his investigation of CAUSE that "[t]he lemma occurs in all genres represented in LOB[7], and in all genres, collocations are predominantly negative" (Stubbs, 1995, p. 5). However, Stubbs (2001, p. 89) analyses UNDERGO and proposes that "people *undergo* serious and unpleasant events, such as medical procedures," and he is careful to mention that an important facet of the lexical profile of UNDERGO is that "[i]n scientific and technical English, the word is usually neutral" (Stubbs 2001, p. 92). Lexical Priming (Hoey, 2005) may help explain Stubbs' observations that the semantic prosody of CAUSE persists across register and genre boundaries, yet UNDERGO appears to be neutral in the scientific registers: "priming is genre and domain specific in the first instance, though there are many primings that apply across generic and domain boundaries" (Hoey, 2005, p. 115). In other words, language users seem to be primed to activate the negative prosody of CAUSE in many or even most genres, but other items, for example *undergo*, appear to be primed to evaluate differently from genre to genre.

In the "Important Note" preceding the body of his book *Lexical Priming: A New Theory of Words and Language*, Hoey writes:

> It can be inferred from the nature of my corpora [the vast majority of which comprises written news texts] that most of the claims I make are to be seen in the first place as restricted to newspaper writing. It will be for others to determine whether they can be extended to other genres or domains or to the language as a whole. (Hoey, 2005, p. vi)

Indeed, Hoey's observations are often qualified with references to the newspaper texts that comprise his corpus. He writes, for example, "[f]or the writers of newspaper text, it would appear that […]" (Hoey, 2005, p. 67); "[…] at least in newspaper English" (Hoey, 2005, p. 86); "[…] in my newspaper data" (Hoey, 2005, p. 90), etc. However, throughout the book, Hoey (2005, p. 13) is also resolute (repeating the claim multiple times) that all ten linguistic features

in his priming hypotheses, including semantic prosody/association, "are in the first place constrained by domain and/or genre. They are claims about the way language is acquired and used in specific situations."

Finally, as we have seen in Chapter 3, Hunston (2007, p. 263) writes, "with respect to CAUSE, […], it would be possible to suggest that this verb loses its association with negative evaluation when it occurs in 'scientific' registers." Hunston (2007, p. 263) does not explore this possibility, however, instead noting immediately:

> A more sustainable argument […] might be that, […], the attitudinal meaning associated with CAUSE applies only when the 'caused entity' concerns animate beings, their activities and goals. Where the 'caused entity' is an inanimate object unrelated to human goals no attitudinal meaning is implied"

Louw and Chateau (2010:261) appear to agree with Hunston on this point. They argue, "In this type of scientific writing, […] the absence of the human context […] removes the need for negativity. The world of hard science is an impersonal world of cause and effect without human agency."

Earlier in the paper, Hunston (2007, pp. 252–253) presents twelve instances of CAUSE from the sub-corpus of the BoE made up of issues of *New Scientist* magazine (hereafter **newsci**) which contain no evidence of the negative semantic prosody of CAUSE. These lines are not intended to demonstrate, however, that the scientific register smooths the prosody, but rather simply to illustrate the more general point that the semantic prosody of CAUSE, while strong, is not absolute; in these twelve examples, "the thing caused is not in itself either desirable or undesirable" (Hunston, 2007, p. 252). That the concordance is constructed from the **newsci** sub-corpus of the BoE appears to be a matter of convenience only.

However, also noted in Chapter 3, Hunston's twelve-line sample comprises only two grammatical patterns of CAUSE, six instances each of **caused by n** and **CAUSE n to-inf**; two examples are reproduced below (see Figure 9.1 for an abridged version of Hunston's twelve-line concordance):

1. minimise the heat loss **caused by convection** while the door is
2. of the tides is like a brake, **causing the spin of the Earth to slow**

No comment is made as to whether these patterns were purposefully chosen or whether the concordance selection was purely coincidental. These patterns, and two more frequently associated with CAUSE, are explored in detail in Chapter 9.

4.5 Conclusion

This chapter has shown that although researchers often leave the terms register and genre undefined, it is possible to clearly distinguish between the two. Biber and Conrad (2009), for example, contrast the defining characteristics of each (see Table 4.1 above) and, furthermore, argue that specific approaches to linguistic features — namely register features, register markers, and genre markers — distinguish the two. Discussion in this chapter has also included a number of linguistic factors that have been found to be affected by register and genre. Finally, it was shown that there does not seem to be any firm agreement as to whether register or genre affect semantic prosody.

Linguistic characteristics, distribution, and interpretation are all important aspects of semantic prosody investigations[8] and these characteristics, as they are defined in Table 4.1, suggest that approaching semantic prosody from the perspective of register is perhaps the more appropriate option of the two. First, instances evincing semantic prosody (e.g. *cause cancer*) are most often not specialized expressions occurring infrequently in a text. Instead, an item is said to have a

positive or negative semantic prosody when a large number of instances of evaluative collocation are observed, usually across many texts in a corpus. Nor is semantic prosody integral to the rhetorical organization of the text. Stubbs (1995, p. 25) has argued that semantic prosody is a kind of cohesive device, providing a continuity of evaluation in a text, but it is difficult to sustain an argument that this specific type of cohesion (via positive/negative semantic prosody) is integral to text construction. Finally, semantic prosodies are, of course, not elements of formatting. Referring again to Table 4.1, it could be argued that instances of semantic prosody are, however, frequent in texts (distribution), and serve important communicative functions (interpretation). Therefore, using Biber and Conrad's defining characteristics of register and genre, a register-specific approach to semantic prosody would seem to be more appropriate than a genre-specific approach.

Notes

[1] The Lancaster-Oslo-Bergen Corpus

[2] http://clu.uni.no/icame/manuals/LOB/INDEX.HTM (accessed 5 October 2017)

[3] The section on linguistic elements of Move 2, Establishing a Niche, also includes brief discussion and examples of "Negation in verb phrases […]; Questions […]; Expressed needs/desires/interests […]; Logical conclusions […]; Contrastive comment […]; and Problem-raising […]" (Swales, 1990, pp. 155–156)

[4] They also discuss "style features", left undiscussed here because aesthetic issues are not directly related to the current research.

[5] Michigan Corpus of Academic Spoken English: https://quod.lib.umich.edu/cgi/c/corpus/corpus?c=micase;page=simple (accessed 6 October 2017).

[6] 'Contagion' is the term used by Bréal (1897, cited in Louw 1993) to refer to this type of "transference of meaning".

[7] LOB refers to the million-word Lancaster-Oslo-Bergen Corpus.

[8] Textual focus is the only characteristic that seems irrelevant in studies of semantic prosody (whether the data is extracted from samples or complete texts does not have any bearing observations of the immediate evaluative environment of the node).

CHAPTER 5:
METHODOLOGY

5.1 Introduction

This chapter discusses the methodological basis for the corpus investigations that follow. The chapter begins in Section 5.2 with a discussion of the items selected for study, including a description of the method employed for revealing multi-word units (MWUs) in corpus data. Section 5.3 discusses the corpora used in the study, and 5.4 looks at the methods of retrieving and analysing data from these corpora. Finally, Section 5.5 outlines the two primary stages in the research, the first quantitative and the second qualitative.

5.2 Items Investigated

5.2.1 Lemmas, word forms, and phrases selected for the study

An important part of the process of deciding which items to investigate in this study was making the decision whether to study lemmas, i.e. all of the inflected forms of a word combined into one 'head word', or to split the study into individual word forms. For example, querying a corpus for the lemma CAUSE returns all instances of the word forms *cause*, *caused*, *causes*, and *causing*. Sinclair (1991, p. 8) argues that "there is a good case for arguing that each distinct form is potentially a unique lexical unit, and that forms should only be conflated into lemmas when their environments show a certain amount and type of similarity." Stubbs (2001c, p. 25) appears to concur: "The word 'word' is ambiguous. First, we have to distinguish between 'lemmas' and 'word-forms'." Stubbs (2001c, p. 27) further argues, that "[…] corpus work provides a lot of evidence that units of meaning are both smaller and larger than the lemma," and for this reason individual word forms ought to be treated separately. Stubbs' (2001c, p. 27) example of CONSUME highlights the potential for overlooking important aspects of an item's

meaning when only a lemma is examined. He shows that there is one literal meaning, "'consume an amount of fuel,'" that is apparently applicable to all of the word forms of CONSUME and all of the word forms share the collocates *more*, *quantities*, *calories*, *energy*, and *oil*. However, a metaphorical meaning of *consuming* is found in the frequent phrases *consuming passion* and *time-consuming*, the latter of which tends to collocate with *costly*, *difficult*, and *expensive*.

The lemma CAUSE was originally selected to act as a 'control' against which the semantic prosodies of other items could be compared. This decision was initially based on the fact that CAUSE has been studied in detail by Michael Stubbs (1995), who shows that collocational evidence for its semantic prosody is very strong; Stubbs (1995, p. 4) reports that 80% of instances of CAUSE studied had negative collocates, only 2% had positive collocates, and the remaining 18% were neutral. Pilot studies of the lemma CAUSE confirmed Stubbs' observations, and further investigations revealed that "the amount and type of evidence" is not notably different for any of the isolated word forms. To illustrate, Table 5.1 below shows the top twenty collocates of CAUSE and its word forms ranked by t-score in the BoE (see Section 2.4.3 for discussion of t-score calculations in corpus data). The table shows there are no particularly noteworthy differences in number or type of evaluative collocate across the word forms that would affect the results of the current study.

However, pilot studies of statistical measures used to determine significant collocates of an item and to observe evidence of semantic prosody in these groups of collocates suggested that such measures, even for an item with a semantic prosody as strong as that of CAUSE, could be problematic.

Table 5.1 Top twenty collocates ranked by t-score of the lemma CAUSE and its word forms in the BoE

	CAUSE (89,830)	*cause* (26,006)	*causes* (9,767)	*caused* (39,790)	*causing* (14,269)
1	by	a	the	by	a
2	a	**problems**	a	a	the
3	the	an	aids	him	**death**
4	**problems**	**serious**	them	the	**problems**
5	him	**cancer**	**problems**	some	them
6	an	**trouble**	**cancer**	**problems**	**concern**
7	some	any	an	an	an
8	them	some	him	her	**trouble**
9	**trouble**	more	me	me	him
10	**serious**	them	us	them	**damage**
11	more	the	some	considerable	some
12	**cancer**	**damage**	more	such	**grievous**
13	**damage**	you	it	**outrage**	serious
14	any	him	**severe**	us	any
15	**concern**	**severe**	such	widespread	bodily
16	me	us	no	**havoc**	more
17	her	such	**pain**	**chaos**	her
18	us	**offence**	people	many	**havoc**
19	**death**	**havoc**	you	great	**severe**
20	**severe**	**chaos**	these	**controversy**	actual

Furthermore, as discussed in detail in Chapters 3 and 9, Susan Hunston uses twelve neutral instances of CAUSE from the BoE sub-corpus of issues of New Scientist magazine to illustrate that the negative semantic prosody CAUSE is not absolute. I observed, however, that only two grammar patterns of CAUSE are represented in Hunston's small concordance. For these reasons, CAUSE became one of the primary focusses of the investigation, and not merely a control item.

The lemma HAPPEN was selected because it, too, is often cited as an example of negative semantic prosody (Sinclair, 1991, 2003; Bublitz, 1996; Partington, 2004, 2014), and yet pilot studies revealed no evidence at all in the collocational profiles to support this claim. Early collocational studies and examinations of concordances suggested that HAPPEN does tend to be found in evaluative environments, but these environments are quite different from those of CAUSE. Put simply, the evaluative nature of CAUSE can be observed directly in its individual

collocates, whereas HAPPEN was observed to evaluate over much larger stretches of text. The only notable difference in the top twenty collocates of HAPPEN and its word forms in the BoE, shown in Table 5.2 below, is that *accident* and *incident*, both collocates of the word form *happened*, stand out as the only evaluative collocates in the entire table. However, it was decided that evidence of negative semantic prosody was too small to warrant further detailed study.

Table 5.2 Top twenty collocates ranked by t-score of the lemma HAPPEN and its word forms in the BoE

	HAPPEN 149,440	*happen* 43,759	*happens* 26,662	*happened* 59,297	*happening* 19,722
1	what	to	what	what	is
2	it	will	it	has	s
3	has	t	that	had	was
4	to	would	this	it	been
5	had	can	whatever	have	be
6	that	could	just	s	this
7	s	not	something	that	are
8	t	things	nothing	never	things
9	will	might	often	just	really
10	this	it	thing	this	it
11	have	never	also	whatever	already
12	would	should	who	nothing	actually
13	whatever	may	never	ever	that
14	things	that	actually	thing	not
15	is	this	anything	**accident**	from
16	never	does	ever	actually	something
17	can	did	which	really	thing
18	just	just	usually	something	were
19	could	you	so	**incident**	all
20	was	ever	always	things	t

It is worth noting that the presence of *accident* and *incident* in the profile of *happened*, however, is arguably due to the journalistic bias in the BoE. In the BNC[1] t-score list, *accident* is the sixty-first collocate of *happened* and *incident* is ranked 108. In the enTenTen13, *accident* and *incident* are ranked 124 and 126 respectively by t-score. Interestingly, however, there are no evaluative collocates ranked higher than these in the t-score lists of *happened* in either corpus. In addition,

because it has been observed to occur as part of many semi-preconstructed phrases (Bublitz, 1996; Partington, 2004), HAPPEN was selected as a candidate for investigations into whether and how phraseological behaviour affects semantic prosody.

A variation of Danielsson's (2007) technique of revealing "multi-word units" (MWUs) by their "cumulative frequency" (Hunston, 2008) was used to identify salient phrases for the four word forms that make up HAPPEN, and this method informed much of the research that followed. As an example, Table 5.3 shows the full development of the phrase ***pronoun*** *don't know what's going to happen* as it 'grows' to the left of the node; the table shows the decreasing frequencies of each "piece" of the MWU as it becomes longer.

Table 5.3 The cumulative frequency of the phrase **pronoun** *don't know what's going to happen* preceded by a personal pronoun

N-8	N-7	N-6	N-5	N-4	N-3	N-2	N-1	Node	Frequency
								happen	43,759
							to	happen	10,938
						going	to	happen	3,758
					s	going	to	happen	1,338
				what	s	going	to	happen	890
			know	what	s	going	to	happen	197
		t	know	what	s	going	to	happen	98
	don	t	know	what	s	going	to	happen	96
I	don	t	know	what	s	going	to	happen	40
we	don	t	know	what	s	going	to	happen	20
you	don	t	know	what	s	going	to	happen	18
just	don	t	know	what	s	going	to	happen	8
they	don	t	know	what	s	going	to	happen	3

Table 5.3 illustrates that the process employed in the current study involves repeatedly calling up concordances and their corresponding Pictures in the BoE (see Section 5.4.2 below for a full explanation of the BoE Picture function). As the table shows, *to* is the most frequent collocate at N-1 for the word *happen*. The concordance of all instances of *to happen* was then called up from within the Picture. A new picture was then generated where it was revealed that *going* is

the most frequent collocate at N-1 of *to happen*. The concordance for *going to happen* was then called up, and a new picture created. This process continued until a salient phrase was 'complete', that is when it was deemed to have reached a 'natural' phraseological or clause boundary[2].

Table 5.4 shows five MWUs for each of the word forms that comprise HAPPEN[3], 'grown' from the five most frequent collocates at N-1. The raw frequencies of some of these MWUs may at first seem too low to warrant investigation (despite there being very little theoretical basis for claiming an absolute cut-off point), but as Hunston (2008:272-273) argues "a sequence that is worthy of note depends on this concept of cumulative frequency rather than on the absolute frequency of the sequence."

Table 5.4 The top-five phrases identified by cumulative frequency for each of the word forms of HAPPEN: happen, happens, happened, happening

Wordform	Phrases based on cumulative frequencies	Raw Freq
happen	[personal pronoun] don't know what's going to happen	81
	[personal pronoun] don't know what will happen	102
	if it doesn't happen	63
	worried about what would happen	12
	the worst thing that can happen	64
happens	[personal pronoun] [will/'ll] have to wait and see what happens	18
	as it happens	1142
	if that happens	532
	if this happens	337
	whatever happens	1369
happened	because of what happened	99
	because of what has happened	26
	because of what had happened	20
	as it happened	578
	who knows what would have happened	28
happening	aware of what is happening	24
	the best of what's happening [on the midpeninsula]	79
	aware of what was happening	42
	what's been happening	204
	what seems to be happening	25

Sinclair (2003, p. 117) notes that many instances of *HAPPEN* are found in environments that express doubt or uncertainty, using the example "I've no idea what will happen" from his sample concordance. Indeed, many of the MWUs in Table 5.4 could be categorized similarly. Pilot studies aimed at identifying the semantic prosodies of these phrases found that evaluative collocates very often occur far beyond the standard span, making the identification and collection of salient data problematic. In addition, *as it happens* and *as it happened* express "the by-chance-meaning of happen" (Bublitz, 1996, p. 17)[4]. Bublitz argues that this meaning of HAPPEN does not evince a negative semantic prosody, and he removes such instances from his own concordances. Therefore, neither of these meaning groups were considered appropriate for the current study.

However, the phrase *the worst thing that can happen* stood out from the set of MWUs and was selected for detailed investigation in part because of the obvious evaluative nature of *worst*. However, although *the best of what's happening* also stood out because of the obvious evaluation provided by *best*, this MWU was ignored because all seventy-nine instances in the BoE were found to be taken from the title of a local newspaper section, "The best of what's happening on the midpeninsula." The process also revealed that *things* is found ninth at N-1 of *happen*, and *these* is the most frequent collocate at N-1 of *things happen*, followed by *make*, *making*, *bad*, and *makes*. The phrases *good things happen*, *bad things happen*, *these things happen* and MAKE *things happen* were therefore selected for study. Similarly, the process revealed that *waiting* is sixth at N-1 of *to happen*, and *accident* and *disaster* are 1st and 2nd at N-1 of *waiting to happen*. The obvious evaluative nature of these phrases made them ideal candidates for this study.

5.3 Corpora Used in This Study

5.3.1 Bank of English

The majority of corpus data presented in this thesis come from the *ca.* 450-million-word Bank of English (BoE) corpus, held by The University of Birmingham, accessed via Telnet. Renouf (1987) provides a detailed description of the process of designing and assembling the corpus.

The BoE was chosen primarily because it is quite large, relatively modern, and was created to be a general reference corpus: "The data of the Bank of English originates mostly from 1990 – 2000 and is intended to reflect the mainstream of current English" (Mahlberg, 2005, p. 42). Indeed, as Renouf (1987, p. 2) writes, the architects of the BoE aimed "to identify those aspects of the English language which were relevant to the needs of the international user." However, it should be noted that the corpus is not especially well-balanced in some potentially important ways. For example, major international varieties of English are not equally represented. Renouf (1987, p. 2) notes that the corpus comprises, "predominantly British English, with some American and other varieties." Figure 5.1, taken from Renouf (1987:3), shows specifically that the designers of the BoE set out to include 70% British English, 20% American English, and 5% "other" varieties. These figures obviously do not add up to 100%, and Li (2015, p. 102) notes that "the remaining 10% represent other types of English," although she does not cite a source for that number.

Figure 5.1 also shows that the corpus is heavily biased toward both male authorship and written English. Renouf (1987:3) offers only that these decisions were made "[f]or different reasons." We might infer what some of these reasons were, though. Renouf (1987, p. 3) notes, for example, that the corpus architects "wished to restrict the choice to works which enjoyed a wide readership." The male:female ratio might then be explained by the unfortunate fact that there

were likely more male authors on the school reading lists, best-seller lists, and in major publisher catalogues that the designers drew upon in the construction of the corpus.

Figure 5.1 General proportions of the BoE, taken from Renouf (1987, p. 3)

book authorship	– 75% male: 25% female
English language variety	– 70% British: 20% American: 5% Other [sic]
language mode	– 75% writing: 25% speech

Additionally, the high written:spoken ratio is likely due to the fact that entering spoken English into the database involved a very time-consuming (and therefore expensive) process of manually transcribing recorded speech. The spoken data is also potentially problematic in yet another way; Walker (2008, pp. 84–85) observes that "well over 50% of the spoken element of the corpus is taken from British and American radio, which is not necessarily representative of natural spoken discourse."

Mahlberg (2005, p. 42) points out, "the Bank of English is sometimes criticised for being too opportunistic because of the heavy reliance on journalistic texts." Indeed, seven of the twenty sub-corpora (refer to Table 4.2) are made up of print news in the form of broadsheets (daily and weekly), tabloids, and magazines comprising a combined total of *ca.* 50% of the total BoE word count. Mahlberg argues, however, that this criticism may be misplaced, because it is possible to argue that "another way to view journalistic texts is to see them as representing mainstream English." Mahlberg's view complements that of Partington (1998, p. 13) — who in turn cites Biber (1988) and Biber et al. (1994) — arguing:

> there is no such thing as 'English as a whole'. All language production belongs to one genre or another, and the English language, like any other, is a collection of genres, none of which deserves the title of 'general English' more than any other. If this is so, then newspaper texts will serve as well as any as the basis of linguistic investigation. […] Moreover, it must be remembered that newspapers consist of not one but a large number of different text types and, in fact, the newspaper section of the corpus is divided into five sections, or mini-corpora […] (home news, arts and features and so on).

In a later article focussing specifically on semantic prosody, however, Partington appears to contradict his own earlier assessment. He writes "I […] decided to concentrate on the corpus of academic writing rather than the newspapers since the latter […] have a tendency to refer drastic and tragic events to their readership" (Partington, 2004, p. 134). This appears to agree with Stubbs (1995, p. 5), who observes that even though the negative semantic prosody of CAUSE persists across the genres of the LOB, "the newspaper press reports have only negative collocations: presumably because newspapers report predominantly crises and disasters!"

Despite such criticisms that the BoE is not balanced, it does, in fact, appear to represent general English as well as can be expected. Data (collocational profiles, MWUs, etc.) taken from the BoE tend to be comparable — at least for the items examined in this thesis — to those taken from the British National Corpus (BNC), well-known for its balanced construction and often used as a reference corpus, and the English Web 2013 (enTenTen13) corpus (Section 5.3.2).

5.3.2 The English Web 2013 corpus

The English Web 2013 corpus (hereafter enTenTen13) is part of the TenTen corpus family. These corpora, accessed via the Sketch Engine[5] website, comprise Internet downloads that have been converted to text only files, then tokenized, deduplicated, lemmatized, and tagged for parts of speech. The minimum target size for the TenTen corpora is ten billion words, and the enTenTen13 is the largest of the English TenTen corpora, comprising over nineteen billion words of running text.

The enTenTen13 was selected for use in the current thesis to compare data observed in the BoE primarily because of its large size. The BNC, by comparison, is generally considered a well-balanced reference corpus, but it has only *ca.* 100 million tokens. This is potentially problematic

in studies of phraseology since larger phrases occur much less frequently in smaller corpora. Sinclair, for example, discusses this precipitous drop in frequency as phrases increase in size. His calculations "suggest that each extra word reduces the number of instances by 83%; this actual number is not important, and the regularity of the reduction in this example is neater than most, but it indicates the scale of the reduction" (Sinclair, 2003, p. 125). Table 5.5 illustrates this problem. The table shows the raw frequencies in the BoE, BNC, and enTenTen13 of phrases analyzed in Chapters 7 and 8 and shows that even in the relatively large BoE, frequencies of some of these phrases are quite low.

Table 5.5 Raw frequencies of the phrases central to investigations in this thesis in the BoE, BNC, and enTenTen13 corpora

	BoE	BNC	enTenTen13
things happen	1,497	352	108,633
these things happen	365	60	9,306
MAKE things happen	366	42	14,935
the worst thing that can happen	64	9	2,300
the best thing that can happen	18	1	602
an accident waiting to happen	66	9	1,428
a disaster waiting to happen	60	6	1,387

Moreover, the BNC, despite claims that it is balanced and representative of general English (if such a thing indeed exists), is simply too small to accommodate the phraseological investigations central to the current study. The single instance of *the best thing that can happen* does not allow for any meaningful comparisons or generalizations to be made.

5.4 Methods of Data Retrieval and Analysis

5.4.1 Lists of collocates

The BoE Lookup software, as accessed via Telnet, allows for the creation of lists of the top fifty collocates of the node word under investigation. The List[6] program makes calculations based

on collocate frequencies within a window of four words to either side of the node[7], often expressed as a "4:4 span" (see Chapter 2 for detailed discussion of span sizes and the implications for corpus research into semantic prosody). Table 5.6 illustrates the 4:4 span for five random lines of CAUSE tagged as a verb in the BoE.

Table 5.6 Five random lines of CAUSE in the BoE, showing the 4:4 span

	N-4	N-3	N-2	N-1	NODE	N+1	N+2	N+3	N+4
1	to	therapy	because	they	cause	pain	to	others	and
2	because	it	does	not	cause	rashes.	Erma	says	that
3	his	divine	mind,	Zeus	caused	the	goddess	to	fall
4	successful	buyer	of	BT	caused	some	movement	in	the
5	years.	The	virus	that	causes	AIDS,	the	Human	Immunodeficiency

The user can choose to rank these lists by raw frequency as collocate (FaC), t-score, or MI score. Table 5.7 above shows the Frequency List, T-List, and MI-List for the top ten collocates of the lemma HAPPEN in the BoE. The collocates are listed without consideration of their position relative to the node; that is, frequencies (FaC) are for all occurrences within the entire 4:4 window.

Table 5.7 BoE-generated collocates Lists (top ten collocates only) for the lemma HAPPEN, ranked by frequency, t-score, and MI score.

	Collocate	FaC	Collocate	FaC	T-score	Collocate	FaC	MI score
1	what	54,612	what	54,612	223.68	iubile	3	8.55
2	to	53,985	it	34,983	130.10	84949	3	8.55
3	the	48,235	that	32,884	114.33	flaura	4	8.23
4	it	34,983	to	53,985	103.63	mohole	3	7.55
5	that	32,884	this	15,470	87.21	icesheets	3	6.81
6	in	24,262	if	10,922	79.39	yeare	3	6.55
7	and	22,196	when	10,038	73.71	midpeninsula	79	6.40
8	is	18,591	t	11,122	73.39	transversions	4	6.38
9	s	17,522	things	5,818	69.38	lamair	3	6.32
10	of	16,890	something	5,794	69.24	impartation	3	6.23

Already apparent in Table 5.7 is that the Frequency and T-score Lists share many of the same high-frequency collocates, but the collocates in the MI-List are very infrequently selected and are distinctive. These lists are discussed in greater detail in Section 6.2.2.

5.4.2 Pictures and positional frequency tables

The BoE Lookup software also gives users the choice of displaying collocates by their position relative to the node. This function is called "Picture" in Lookup and is sometimes referred to as a "positional frequency table" (hereafter PFT) (*cf*. Stubbs 2001b, p. 94). The Picture creates separate lists according to their position relative to the node rather than basing calculations on frequencies within the whole 4:4 span. Table 5.8 shows an abridged Frequency Picture (showing only the top-twenty collocates) of HAPPEN in the BoE. The table shows, for example, that *what* is the most frequent collocate in three of the four positions to the left of the node, but at N-3 the most frequent collocate is *that*.

Table 5.8 Abridged frequency picture of the 4:4 span of HAPPEN in the BoE; abridged, showing only the top twenty collocates in each column

	N-4	N-3	N-2	N-1	NODE	N+1	N+2	N+3	N+4
1	what	that	what	what	NODE	to	the	i	i
2	that	it	it	it	NODE	in	be	time	t
3	t	what	that	has	NODE	when	you	past	it
4	something	this	going	to	NODE	if	me	t	you
5	it	know	this	had	NODE	at	i	it	when
6	but	something	if	that	NODE	<p>	that	was	we
7	this	about	something	s	NODE	again	them	when	happen
8	if	if	things	t	NODE	i	him	we	is
9	know	things	see	will	NODE	but	we	ago	not
10	think	thing	when	this	NODE	here	us	you	happened
11	tell	t	nothing	have	NODE	next	my	next	that
12	thing	but	as	would	NODE	on	her	don	know
13	things	see	thing	whatever	NODE	now	it	people	was
14	don	is	doesn	things	NODE	before	this	last	ago
15	i	s	anything	is	NODE	because	quickly	if	think
16	nothing	when	about	never	NODE	so	your	did	few
17	not	believe	know	can	NODE	after	there	is	if
18	never	think	didn	just	NODE	there	our	future	there
19	best	how	could	could	NODE	with	said	there	time
20	find	just	did	was	NODE	during	know	country	don

Unfortunately, Sketch Engine, which holds the version of the enTenTen13 used in this thesis does not offer Picture creation software or similar Positional Frequency Table (PFT) options. Therefore, it was often necessary to create PFTs in Microsoft Excel from downloaded or copied

concordances. All references to "PFT" in this thesis refer to such bespoke tables. These PFTs are created by first copying concordances into Excel. All capital letters are converted to lower case, and punctuation and tags are removed from the concordances. Then, Excel's built-in "text to columns" function is used to isolate each word in its own cell, and empty cells are removed. Columns are then shifted as required to centre the node and align the words in their correct positions relative to the node. Copies of each individual column in the 4:4 span are pasted to empty columns on the page, where duplicate items are removed. The remaining unique items in this list are used as references to count the instances in the concordance using Excel's "countif" formula. Once counted, each column is sorted from highest to lowest frequency. The result is a positional frequency table.

Creating Positional T-score Tables in Excel was not considered feasible during the data preparation stage of analysis because (as explained in detail in Chapter 2), t-score calculations require the collocate's raw frequency in the whole corpus for comparison to its specific frequency in the concordance. Not only was I unable to locate an accurate frequency list of all tokens in the 450-million-word version of the BoE to accommodate such calculations, it is questionable that Excel and the computer used to create these tables have the processing power needed to make the many hundreds of consecutive calculations required.

5.4.3 Concordances: sample sizes and line lengths

Because both CAUSE and HAPPEN are very high-frequency verbs, sample concordances of *ca.* 500 random lines from the BoE were used for qualitative analyses in the chapters that follow. The nature of the random selection of lines in the BoE is worth mentioning, as this method contributes to the replicability of the studies presented. The Lookup software creates what are more accurately termed pseudo-random concordances by selecting every n^{th} instance from the

total occurrences: "For example, if there are 1,000 instances of a given word, and the search request specifies 100 examples, the software will take display [sic] every tenth occurrence" (Hunston, 2011, p. 8).

Some of the phrases and grammatical patterns analysed here are relatively rare (**CAUSE n n**[8], for example) and some sub-corpora of the BoE are relatively small, however, so it was not always possible to collect 500 instances. Even when it was possible to collect 500 line-concordances they often ended up being edited significantly, as mis-tagged and otherwise irrelevant instances were removed before analysis began (these concordance edits are discussed in detail in the relevant sections). Furthermore, 'full' sample concordances were often divided into shorter sets when different senses of the phrases were observed (again, discussed in the relevant sections below).

For analysis of grammatical patterning multiple concordances had to be combined in every case. This is mainly because the BoE query language makes searching for some patterns problematic. For example, **CAUSE n** is both a pattern in its own right (*smoking **causes cancer***) and is also 'embedded' in other patterns examined, for example the ditransitive **CAUSE n n** (*caused him a lot of problems*), and **CAUSE n to-inf** (*cause arteries to clog*). This means that any instances of **CAUSE n** in the BoE must be extracted manually from the concordances. Figure 5.2 illustrates how multiple patterns can returned from a single corpus query. Of these lines, five (4, 5, 12, 13, 14) represent the **CAUSE n to-inf** pattern, and line 10 appears to be a passive voice variant of the **CAUSE n n** pattern: *a part of me that's guilty for the amount of pain I've caused women*. This leaves fourteen examples of the **CAUSE n** pattern of the original twenty lines.

Figure 5.2 Twenty random BoE lines of CAUSE tagged as a verb followed immediately by a noun

1. **cause** pain to others and have become so numbed to the
2. **cause** world famine, world chaos and probably world war; and
3. **cause** problems if you are driving in a strange city
4. **cause** leaves to turn yellow-brown and brittle. To get rid of
5. **cause** investment, jobs and output to overshoot as well. It is
6. **cause** confusion among staff trying to glean details for
7. **cause** abnormalities in sperm: why take the chance? One other
8. **cause** disruption. Labour will not quickly forget how it lost

9. **caused** confusion and anger among many Israelis. For National
10. **caused** women. I can understand why Cheryl feels she should see
11. **caused** resentment within his party. But, if he can bring
12. **caused** Centrelink to overpay Newstart recipients by between $2
13. **caused** San Juan to be known throughout the Spanish-speaking

14. **causes** Atlantic depressions to track northwards across the
15. **causes** AIDS, within the school system. <p> AIDS has killed 251
16. **causes** damage to a lamp near the window, the lamp damage would

17. **causing** meningitis. Royce Johnson says that there's no wonder
18. **causing** feelings of uncertainty and unfairness. On the other
19. **causing** controversy. The local building industry says it could
20. **Causing** fear. scenario: Imagined scene. nonchalantly:

A related difficulty is that the patterns refer to noun phrases not just individual, isolated, nouns. The BoE, however, is not capable of searching for noun phrases automatically. Multiple queries were employed, however, to ensure that the patterns examined included examples of multi-word noun groups (i.e. including adjective pre-modifiers). In Chapter 9, for example, seven separate queries returned a total of 528 lines of the **CAUSE n n** pattern in **newsci**, but only thirteen of these lines actually evinced the pattern. Not all possible constructions were exhausted in this process, but these multiple queries resulted in what appear to be concordances generally representative of the patterns (Chapter 9, specifically Endnote 3, provides more details on the queries employed).

The BoE Lookup software accessed via Telnet defaults to a concordance line length of eighty

characters. This is potentially problematic for a number of reasons directly related to this study. First, long phrases like *an accident waiting to happen* (twenty-nine characters, including spaces) and *the worst thing that can happen* (thirty-one characters, including spaces) can occupy a sizable proportion of the default eighty-characters, leaving very little in the environment to analyse. Second, the research presented here necessarily requires extended contexts for evidence of semantic prosody and textual evaluation (see Section 2.6).

The BoE software allows users to save files, and in the process of saving users are prompted to select the line length, in number of characters, they require. These lines can then be emailed directly from the BoE server to the user. Unfortunately, approximately halfway through the preparation of data for this thesis, the email function of the BoE software ceased working for me, and I was forced to manually copy and paste concordances directly from the Telnet window. In order to copy lines longer than the 80-character default, the Telnet settings had to be adjusted. For much of the research presented here, the goal was to analyse lines of 200 characters each. This number is somewhat arbitrary but was meant to allow for a span of at very least 10:10 (and in most cases, it was much more). Due to idiosyncrasies of the settings in the Telnet program used, however, many of the lines turned out to be somewhat shorter, (*ca.* 180 characters). Furthermore, concordances presented here are often 'trimmed' of unnecessary items at the beginnings and ends of lines. This was done to improve readability by simplifying formatting (shorter lines align more easily on the page making relevant data more easily observable) and cutting away the excess 'noise' from the pertinent data, saving the reader time and effort.

5.4.4 A note on CAUSE tagged as a verb in the BoE

This section briefly discusses an idiosyncrasy of the BoE Lookup software that affects analyses

of CAUSE. Specifically, two different raw frequencies are reported by the BoE for the lemma CAUSE tagged as a verb[9]. Table 5.9 shows the frequencies returned for a sample of related queries in the BoE. The queries in the table are presented as they are made in the Lookup software: The "@" symbol indicates the lemma is requested and appending the query with "/VERB" requests only instances tagged as a verb in the corpus. The first column of the table, therefore, shows results for the lemma CAUSE tagged as a verb; the second column shows results for the lemma CAUSE tagged as a noun (i.e. the frequency includes 36,952 instances of CAUSE tagged as a noun); the third column shows frequencies of the lemma CAUSE with no tag specified; and finally, the fourth column shows the frequencies of the word form *cause* tagged as a verb. Table 5.9 shows that the first time the corpus is interrogated for CAUSE tagged as a verb, the software returns 98,259 instances; a second identical interrogation, however, returns a frequency of 89,830. This is a discrepancy of 8.6%. A third query returns the same frequency as the first, and so on; results appear to alternate between these two numbers.

Table 5.9 Frequencies of the lemma CAUSE tagged as a verb, as a noun, tag unspecified, and the word form *cause* tagged as a verb in the BoE

	cause@/VERB	cause@/NOUN	cause@	cause/VERB
1st query freq	98,259	36,952	126,782	26,006
2nd query freq	89,830	36,952	126,782	26,006
Difference	8,429	0	0	0
% Difference	8.6%	—	—	—

The same discrepancy is observed for interrogations of the lemma CAUSE tagged as a verb in all the sub-corpora, e.g. **newsci** returns 4,969 and 4,682 occurrences of cause@/VERB, which is a 5.8% difference. The table also shows that that this error only appears to affect the cause@/VERB (lemma plus tag query), although I have only checked CAUSE and HAPPEN. That is, as the table above shows, cause@ is unaffected, as is cause/VERB, and so on.

There are two ways to deduce the 'correct' frequency of CAUSE in the BoE. First, since CAUSE tagged as a noun and CAUSE (with no tag specified) are consistently reported at 36,952 and 126,782 instances respectively, we can simply subtract the noun instances from the total and we have 89,830, which matches the second query report. Another approach is simply to sum the frequencies reported for the word forms, since they are reported consistently and, therefore, assumed to be correct. Again, we see a total of 89,830, shown in Table 5.10

Table 5.10 Frequencies for each of the word forms of CAUSE in the BoE: the total is equal to cause@/VERB returned by the 'second query' of the BoE

BoE Query	Frequency
cause/VERB	26,006
causes/VERB	9,765
caused/VERB	39,790
causing/VERB	14,269
Total	89,830

This frequency discrepancy is noted here mainly because CAUSE is central to the analyses and arguments presented in this thesis, and reproducibility is an important facet of the thesis. The difference between the two reported frequencies is large enough to potentially create confusion for anyone attempting to replicate the results presented here using the BoE. Therefore, it is important for the reader to be aware that in this thesis all profiles and concordances of CAUSE tagged as a verb are taken from the second query, which returns 89,830 lines.

5.5 Stages in the Research

5.5.1 Stage one: quantitative collocational analyses

The first stage of the research involved quantitative collocational analysis. This was accomplished by creating collocational profiles using BoE Pictures and my own PFTs. For these analyses, the top fifty collocates (unless otherwise noted) in the 4:4 span were examined. BoE collocates were ranked by t-score, and data in the PFTs were ranked by frequency. In most

cases, t-score and frequency data are quite similar and there did not appear to be any noteworthy discrepancies in the data sets. In fact, for many of the profiles the t-scores themselves were not considered (unless they fell below the range considered significant, as discussed in Chapter 7), and raw frequency data (FaC) were often collected and used as well.

Further, levels of statistical significance — either in comparisons of the frequencies of positive and negative collocates or of register variation — are not discussed. That is, no attempt to generalize using inferential statistical methods is made in the analyses that follow. This decision was made largely for reasons cited at length above (see especially Section 2.4.4), primarily that language is never random, and these inferential methods rely principally on the foundation that observed frequencies can be compared to expected frequencies in a random data set.

Positive and negative words in the Picture/PFT profiles and concordances were labelled and counted using Microsoft Excel. This semi-automatic process was designed to alleviate the extremely time-consuming and tedious process of manually scanning and labelling thousands of concordance lines and scores of Lists, Pictures, and PFTs. The method employed here is inspired by Dilts and Newman (2006), who sought to establish lists of evaluative collocates "prior to, and independently of, corpus-based studies of prosody" (Dilts and Newman, 2006, p. 240). However, as the following account of the procedure illustrates, both the data and lists were also manually checked and updated, *post hoc*, to ensure that no evaluative collocates were missed. This created a circular process: the lists assisted in the labelling of the data, and the data facilitated expansion and 'fine-tuning' of the lists.

To begin, rudimentary lists of negative and positive words were downloaded from the Internet[10]. These lists required a great deal of editing and preparation, however, before they were

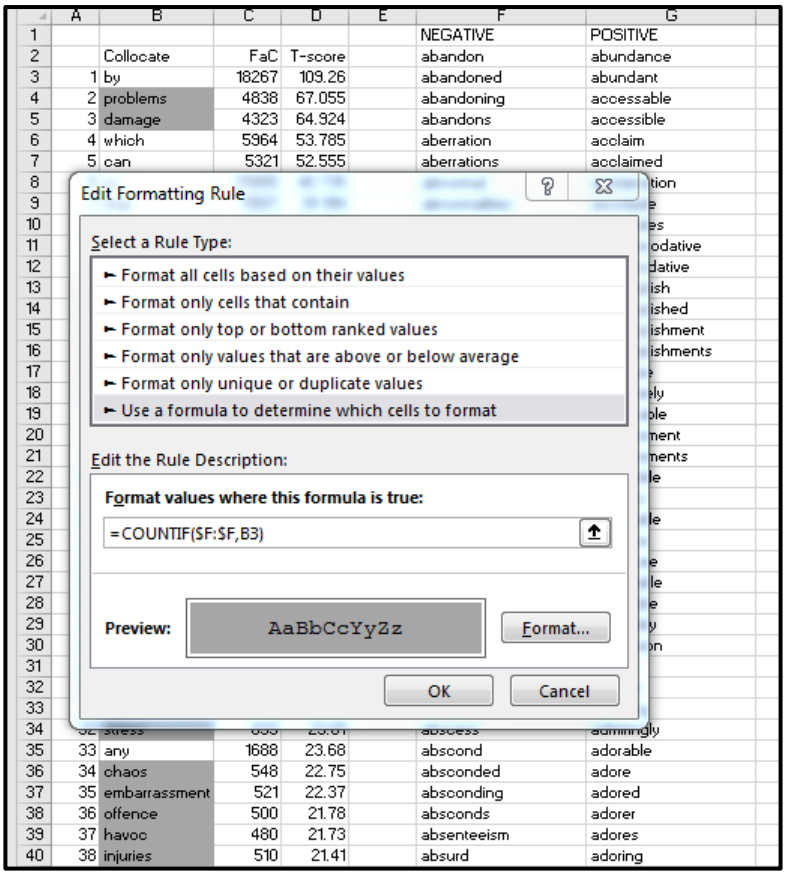
considered usable. For example, the original lists comprise individual word forms only, or at times incomplete sets of related word forms; for example, the original list contains the noun forms *accusation* and *accusations* and the verb forms *accuse*, *accuses*, *accusing*, *accusingly*, but not *accused*, which was added manually.

Once edited, these lists were copied into Microsoft Excel and the "conditional highlighting" function was used to identify matching word forms in the corpus data; positive and negative word forms in a concordance, List, Picture, or PFT were coloured green or red respectively. After the initial matching and colour-coding stage of analysis was completed, the data was carefully double checked. Inappropriate matches were removed from the lists or additional words included as noticed in the data. Conditional formatting was then replaced with static formatting, made possible by the Excel software add-on ASAP Utilities[11], so that automatic counting by cell colour could take place. The screen capture in Figure 5.3 below shows a section of an Excel spreadsheet containing the BoE collocates List for CAUSE. The formula in the window refers only to the B3 cell; the "formula painter" was used to copy this formatting rule to the rest of the cells in the B column (which, as the capture shows, has been done in advance for illustrative purposes here).

The lists of positive and negative collocates are in a state of continuous development and are not meant to be considered exhaustive. At the time of writing they held *ca.* 2,000 positive word forms and *ca.* 8,100 negative word forms, and the process of adding and deleting word forms will continue as they are noticed in data. Their coverage at this point, though, is strikingly comprehensive, and the point of using such lists in the current study is not, as Dilts and Newman propose, "to explore prosody with less dependence on a researcher's subjective, evaluative judgments" but simply as a time-saving method of observing and amassing groups of evaluative

collocates quickly and relatively easily.

Figure 5.3 Screen capture of a portion of a Microsoft Excel spreadsheet and conditional formatting rule window; the formula identifies and shades collocates in the B column if they match words in the F column.



	A	B	C	D	E	F	G
1						NEGATIVE	POSITIVE
2		Collocate	FaC	T-score		abandon	abundance
3	1	by	18267	109.26		abandoned	abundant
4	2	problems	4838	67.055		abandoning	accessable
5	3	damage	4323	64.924		abandons	accessible
6	4	which	5964	53.785		aberration	acclaim
7	5	can	5321	52.555		aberrations	acclaimed
35	33	any	1688	23.68		abscond	adorable
36	34	chaos	548	22.75		absconded	adore
37	35	embarrassment	521	22.37		absconding	adored
38	36	offence	500	21.78		absconds	adorer
39	37	havoc	480	21.73		absenteeism	adores
40	38	injuries	510	21.41		absurd	adoring

This method can provide, at a glance, an impression of the prosody of an item (or lack thereof), that, once established, can be investigated in more detail. However, it remains that "the labels that are assigned […] are necessarily the analyst's – it is s/he who decides how to interpret, categorize, and classify the collocates semantically" (Bednarek, 2008, p. 122). In addition, no claim is made here that the collocates identified as "good" or "bad" are objectively so. As Dilts and Newman (2006, p. 234) argue, "[i]t may be the case that the study of prosody is always likely to involve a certain degree of subjectivity on the part of the researcher." I have made decisions on words to add to and omit from the lists of "positive" and "negative" collocates that

other researchers might disagree with. For example, I omitted *cheap, cheaper, cheapest,* and *cheaply* from the negative list despite their occasional negative association[12]. However, *cheapen* and related word forms were left in the list because of their clearly negative core meanings. These minor edits notwithstanding, this process works quite well. Because semantic prosody relies on the observation of a preponderance of positive or negative collocates, disagreements regarding the labelling of a small number of word forms in a list or picture will not usually affect the overall observation.

5.5.2 Stage two: qualitative examination of concordances for comparison

The second stage of the study involved qualitative analysis of sample concordances (see above). This analysis involved close readings of the same ca. 200-character concordances used for the quantitative analysis described above. Lines were read carefully, and any indications of evaluation were highlighted. Lines were then sorted and categorized by the type of evaluation observed, i.e. prosodic or textual (see Section 2.6).

Qualitative analysis was deemed necessary in the current study for at least three reasons. First, close reading of the lines is required to reveal evaluative phrases acting as collocates of the node because Lists, Pictures, and PFTs contain single-word collocates only. Secondly, qualitative examination was also required to find collocates and phrases occurring outside of the standard 4:4 span. For example, *fear* is found at N-5 in line 21, and *dread* occurs at N-6 in line 22. Although the BoE Picture output can be set for a 6:6 window, which would capture both *fear* and *dread* in these lines, many evaluative collocates are found in larger contexts, as analyses that follow will show. Finally, even when evaluative words and phrases are identified, there remains the question of whether they constitute evidence of semantic prosody or textual

evaluation, which is also observable only in qualitative examination of concordances (and, at times, requires much larger stretches of text than a concordance line). In the case of *fear* and *dread*, for example, neither appear to be have a strong syntactic relationship with the node.

21. workers in Kukes fear that the opposite is **happening**. Nato is
22. `I dread to think what might have **happened** if she hadn't been

In line 23 below, there are no single-word collocates that could be identified as evaluative in a collocational profile. However, the phrases *greenhouse gases* and *global warming* both evaluate negatively despite comprising words that, on their own, are neutral. While the single word collocates *greenhouse* and *global* in the collocational profiles suggest as much, qualitative scrutiny is required to confirm.

23. Particularly as carbon dioxide one of the greenhouse gases **causing** global warming, is produced in the main by

Similarly, qualitative examination allows for the identification of much longer phrases or even whole clauses that evaluate. For example, there are no negative single-word collocates in line 24, but the clause *my body feels out of control* evaluates negatively.

24. `Is this me? Is it not me? What's **happening** to me? My body feels out of control

However, this instance would not be considered an instantiation of semantic prosody. For reasons discussed in detail in sections 2.6 and 2.7, this line is considered to evoke textual evaluation.

5.6 Summary

This chapter has discussed the methodology underlying the studies presented in this thesis. First, the chapter outlined the items investigated. The corpora used and the types of data

(concordances, collocates Lists, Pictures, and PFTs) extracted from these corpora have been discussed in detail. This involved discussion of semi-automatic methods of identifying positive and negative collocates and of revealing MWUs in corpus data. Further, the anomalous reporting of the frequency of the lemma CAUSE tagged as a verb in the Bank of English has been noted. Finally, the two primary stages of research, quantitative and qualitative, into the lemmas and their phraseological behaviour has been presented.

Notes

[1] The British National Corpus (BNC), accessed via www.sketchengine.co.uk, is a balanced corpus of 100-million words of spoken and written English. The BNC is not one of the primary corpora used in this study, but it was used periodically to make very general comparisons.

[2] Danielsson's (2007, p. 19) method is quite different. Most significantly, she uses the most frequent collocates in the whole 4:4 span, regardless of position, and she discards grammatical/function words (at least in the early part of the process). Danielsson also sets an arbitrary cut-off of five occurrences.

[3] Of course, many of these MWUs could be extended to the right of the node. The phrase *the best of what's happening on the midpeninsula*, for example, was identified because my intuition suggested that *the best of what's happening* was over-represented in the corpus. It was discovered that all instances of this MWU are from the title of a local newspaper section.

[4] The article, "Semantic prosody and cohesive company: somewhat predictable" was obtained through personal communication with Professor Bublitz. Page numbers refer to this personal copy and not to those in the original publication, which I was not able to access.

[5] https//:www.sketchengine.co.uk

[6] Capitalization of List and Picture in this thesis refers to the Lookup software output.

[7] The node is the word or phrase under investigation. The notations N+ (plus) or N- (minus) a number refer to the positions around the node.

[8] **CAUSE n n** represents the ditransitive uses of cause, i.e. cause followed by two noun groups, as in *it would **cause me concern***.

[9] The phrase "tagged as a verb" is used purposefully throughout this thesis to reflect that in it is not, strictly speaking, correct to refer to a lemma or word form "as a verb" in corpus data. There are many instances in the BoE, for example, of the count nouns *cause* and *causes* tagged

incorrectly as verbs. Concordance analyses presented in this thesis included as an initial step removing mis-tagged instances.

[10] http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html; the original positive list is accessible at: http://ptrckprry.com/course/ssd/data/positive-words.txt; the original negative list is accessible at: http://ptrckprry.com/course/ssd/data/negative-words.txt; (all pages last accessed 19 October 2017).

[11] See: http://www.asap-utilities.com/index.php

[12] The second usage listed in the CCED (2001:248): "If you describe goods as **cheap**, you mean they cost less money than similar products but their quality is poor."

CHAPTER 6:
POTENTIAL DIFFICULTIES OBSERVING COLLOCATIONAL EVIDENCE OF SEMANTIC PROSODY

6.1 Introduction

As illustrated in Chapter 2, the meanings of "collocate", and "collocation" are variable and are not necessarily consistently applied across studies of semantic prosody. It is generally uncontroversial to assert that semantic prosody is collocational (Barnbrook, Mason, and Krishnamurthy 2013; Bednarek 2008; Bublitz 1996; Louw 1993, 2000; McEnery, Xiao, and Tono 2006; Morley and Partington 2009; Partington 2004; Stubbs 1995; Walker 2011a, 2011b; Xiao and McEnery 2006), but defining it thus necessarily begs the serious question: which sense of "collocate" and "collocation" are being employed in the analyses?

This chapter addresses some of the potentially problematic aspects of collocational evidence of semantic prosody. The chapter is divided into four main sections. First, 6.2 looks at how collocational profiles can be problematic in studies of semantic prosody. Corpus data show that both the method of observing the collocates (in sorted concordance lines, ranked lists, or positional frequency tables) and the measure of their statistical significance (The BoE allows for analysis based on raw frequency, t-score, or MI) have potentially strong effects on claims that an item has a semantic prosody. The section concludes with a discussion of how total FaC values of evaluative collocates may be employed in interpretations of corpus data; this method is central to analyses of phraseological behaviour and grammatical patterning in the Chapters that follow.

It is argued in Section 6.3 that even when statistical measures are abandoned, the simple co-occurrence aspect of collocation can be problematic in identifying an item's semantic prosody.

Corpus data illustrate that the often-complex relationship between node and collocate can have substantial effects on observations of an item's mode of evaluation (see Section 2.6), and its polarity. Finally, Section 6.4 presents the results of qualitative analyses of concordances of CAUSE and HAPPEN.

6.2 Potential Problems Observing Semantic Prosody in Collocational Profiles

This section discusses potential problems encountered when attempting to observe semantic prosody in groups of statistically relevant collocates. McEnery and Hardie (2012, p. 127) argue: "Because the analyst's choice of statistic has such a major effect on the outcome, there is in effect an inherent subjectivity in the determination of what is, and what is not, a collocate." Therefore, there is a potentially equal subjectivity in claims that an item has a positive or negative semantic prosody.

6.2.1 Quantitative concordance analysis and the notion of span

In their discussion of collocational significance, Barnbrook et al. (2013, p. 79) argue: "Mostly the only choice one has is to vary the significance measure, but choosing the right span is probably even more important." The location of a collocate relative to the node word, i.e. the span within which that collocate occurs or the precise location at which a collocate is found, is a critical aspect of any calculation of statistical significance. This section discusses a relatively simple method of quantitative collocational analysis, namely the semi-automatic identification of evaluative collocates in a KWIC concordance (see Section 5.5.1 for details on the methods used to identify positive and negative words in the concordance lines).

The method employed in this section is akin to the "collocation-via-concordance approach" (McEnery and Hardie, 2012, p. 126). The investigation focusses on two spans, the standard 4:4,

and a much larger 10:10 in concordances of 500 random lines of CAUSE and HAPPEN in the BoE. Before the examination began, fifty-two instances of CAUSE as a noun mistakenly tagged as a verb, as well as occurrences of the informal conjunction *'cause* were removed from the concordance, leaving 448 lines for analysis. Results are shown in Table 6.1.

Table 6.1 Identification of evaluative collocates in spans of 4:4 and 10:10 for 500 random BoE lines of CAUSE

	Freq.	%
Lines with at least one negative collocate in the 4:4 span	321	71.5%
Lines with at least one negative collocate in the 10:10 span	381	84.9%
Lines with at least one positive collocate in the 4:4 span	44	9.8%
Lines with at least one positive collocate in the 10:10 span	103	22.9%
Lines with no evaluative collocates in the 4:4 span	110	24.5%
Lines with no evaluative collocates in the 10:10 span	49	10.9%

As Table 6.1 shows, the negative semantic prosody of CAUSE is readily apparent in collocates observed in the concordance: 321 of the 448 (71.5%) lines have at least one negative collocate in the 4:4 span, and 383 (84.9%) have at least one negative collocate in the 10:10 span. In comparison, only forty-four lines (9.8%) have a positive collocate in the 4:4 span, and 103 lines have a positive collocate in the 10:10 span. Not shown in the table is the fact that positive and negative collocation are not mutually exclusive; twenty-eight lines have at least one negative and one positive collocate in the 4:4 span, and 103 lines have at least one negative and one positive collocate in the 10:10 span. Also notable is that 110 (24.5%) lines have no evaluative collocates in the standard span and forty-nine lines (10.9%) have none in a span of 10:10. It will be demonstrated in sections that follow, however, that many of the lines with no evaluative collocates in fact contain lexical phases comprising two or more neutral words that only together are observed to evaluate (see 6.3 and 6.4 for detailed discussions of these structurally complex collocates).

Results of the semi-automatic identification of evaluative collocates in 500 random lines of HAPPEN from the BoE (no lines were removed) are shown in Table 6.2.

Table 6.2 Identification of evaluative collocates in spans of 4:4 and 10:10 for 500 random BoE lines of HAPPEN

	Freq	%
Lines with at least one negative collocate in the 4:4 span	98	19.6%
Lines with at least one negative collocate in the 10:10 span	212	42.4%
Lines with at least one positive collocate in the 4:4 span	68	13.6%
Lines with at least one positive collocate in the 10:10 span	166	33.2%
Lines with no evaluative collocates in the 4:4 span	347	69.4%
Lines with no evaluative collocates in the 10:10 span	184	36.8%

As Table 6.2 shows, in the 4:4 span, only ninety-eight (19.6%) of the lines contain negative collocates, sixty-eight (13.6%) contain positive collocates, and 347 (69.4%) contain no evaluative collocates at all. In all, 184 lines (36.8%) have no evaluative collocates even in the 10:10 span. Again, this is due at least partially to occurrences of structurally complex collocates in the concordance, but as demonstrated in Sections 6.3.2, these instances are often indicative of a mode of evaluation that is not prosodic.

Corpus data in the form presented in this section show that the notion of span as it relates to semantic prosody can be problematic. On the one hand, the extended 10:10 span is not required to reveal a convincing number of negative collocates of CAUSE; more than enough relevant evaluative collocates are found in the 4:4 window to support the argument that CAUSE has a negative semantic prosody. Nor does the larger span reveal counter-evidence in the form of positive collocates that might challenge the notion the CAUSE has a negative semantic prosody. In the case of HAPPEN, however, the extended span does reveal many additional negative collocates. Whether a total of 42.4% is enough to support a convincing argument for negative semantic prosody is debatable, however. Further, the status of these negative "collocates" in

the extended span remains questionable (i.e. whether they exhibit logical relationships with the node, etc.). Therefore, further research is needed before it can be claimed that these collocates represent evidence of the semantic prosody of HAPPEN.

6.2.2 BoE-generated T-lists of collocates of CAUSE and HAPPEN

The notion of span is also central to statistical calculations of collocational significance. This section looks at one such method, namely automatically generated T-Lists of collocates in the 4:4 span. Table 6.3 shows the top fifty collocates in the BoE-List for CAUSE tagged as a verb.

Table 6.3: BoE-generated T-list of collocates of the lemma CAUSE in the 4:4 span, negative collocates in bold

	Collocate	FAC	T-score		Collocate	FAC	T-score
1	by	18,497	107.86	26	**injury**	783	25.75
2	**problems**	4,931	67.51	27	or	3,970	25.64
3	**damage**	4,351	65.06	28	**stress**	708	25.22
4	which	6,200	53.59	29	this	4,753	24.85
5	can	5,465	51.92	30	of	24,107	24.78
6	to	27,588	47.71	31	stir	631	24.64
7	that	12,061	37.08	32	**virus**	643	24.49
8	may	2,852	36.94	33	**severe**	654	24.43
9	**cancer**	1,486	36.86	34	any	1,752	23.18
10	**death**	1,613	35.99	35	**chaos**	551	22.76
11	**disease**	1,395	35.69	36	such	1,566	22.62
12	**trouble**	1,421	35.68	37	**embarrassment**	524	22.40
13	could	2,908	35.12	38	**offence**	500	21.74
14	**pain**	1,312	34.61	39	**havoc**	480	21.72
15	**harm**	1,111	32.77	40	**injuries**	519	21.52
16	the	50,687	32.23	41	bodily	462	21.24
17	**concern**	1,150	31.96	42	**disruption**	452	21.04
18	has	4,615	28.72	43	**confusion**	476	21.00
19	serious	1,004	28.29	44	**anxiety**	462	20.66
20	**loss**	931	27.51	45	likely	724	20.66
21	some	2,500	27.39	46	blood	589	20.61
22	among	1,095	26.83	47	**suffering**	504	20.53
23	**problem**	1,085	26.47	48	effects	535	20.53
24	**deaths**	730	26.11	49	**symptoms**	454	20.26
25	**distress**	693	25.98	50	much	1,425	20.23

Of the fifty collocates listed by the BoE software, seventeen are closed-class grammatical words; of the remaining thirty-three lexical words, twenty-seven are negative, and six are neutral (*serious*, *stir*, *bodily*, *likely*, *blood*, *effects*).

In contrast, Table 6.4 below shows that none of the collocates in the T-List of HAPPEN in the BoE are negative. This is somewhat surprising because the collocation-via-concordance approach above revealed that almost 20% (98 of 500) of the lines contained a negative collocate in the 4:4 span.

Table 6.4: BoE-generated T-list of collocates of the lemma HAPPEN in the 4:4 span

	Collocate	FAC	T-score		Collocate	FAC	T-score
1	what	54,612	223.68	26	s	17,522	45.61
2	it	34,983	130.10	27	could	4637	45.45
3	that	32,884	114.34	28	again	3057	45.25
4	to	53,985	103.64	29	see	3586	43.95
5	this	15,470	87.21	30	me	4,094	42.56
6	if	10,922	79.39	31	anything	2,267	41.29
7	when	10,038	73.71	32	not	8107	40.53
8	t	11,122	73.39	33	can	5150	37.28
9	things	5,818	69.38	34	might	2260	35.28
10	something	5794	69.24	35	was	12,559	34.84
11	i	16,504	58.26	36	did	2691	34.43
12	thing	3923	56.39	37	really	2,138	33.63
13	nothing	3767	55.91	38	so	4,793	33.22
14	had	9067	54.81	39	doesn	1531	33.15
15	going	4467	54.76	40	here	2,246	32.97
16	has	9272	54.21	41	think	2709	32.48
17	is	18,591	54.10	42	next	2,177	31.92
18	know	4,737	51.70	43	actually	1,414	31.09
19	will	7,820	50.05	44	we	6,428	30.76
20	whatever	2,738	49.91	45	exactly	1151	30.67
21	just	5,160	49.50	46	have	8,265	30.64
22	never	3603	49.04	47	all	5198	30.63
23	but	10,669	48.56	48	because	2946	30.53
24	would	6469	48.03	49	why	1781	30.37
25	you	10,921	47.16	50	ever	1,581	30.33

The complete absence of negative collocates in the BoE T-List is also surprising given that at least three well-known researchers have studied the semantic prosody of HAPPEN — Bublitz (1996), Partington (2004; 2014), and Sinclair (1991; 2003) — and these studies, especially Sinclair's, are very commonly cited in the literature where HAPPEN is provided as a well-known example of negative semantic prosody (Louw, 1993, p. 158; Stubbs, 1995, p. 3; Cotterill, 2001, p. 292; Tognini-Bonelli, 2001, p. 111; Whitsitt, 2005, p. 287; Xiao and McEnery, 2006, p. 106; Bednarek, 2008, p. 121; Granger and Paquot, 2008, p. 31; Kennedy, 2008, p. 36; Ellis, Frey and

Jalkanen, 2009, p. 90; Stewart, 2010). However, Bublitz (1996, p. 14) does make the important observation that, "[t]he semantic profile of *happen* remains obscure due to unclear reference. The grammatical subject of *happen* is either outside the given span or, when inside, it is a semantically unspecific pronoun (*it, that*), an anaphoric or general noun (*thing*)." Recall that the "given span" used by the BoE to create the collocates Lists is the standard 4:4, and is static in the creation of Lists, unlike the Picture software, which allows the investigator to choose a span from 3:3 to 6:6 (which is likely still too small to capture the semantic prosody of HAPPEN).

Data in Table 6.4 above supports Bublitz's argument. The general nouns *things* and *thing*, central to phrases analysed in Chapter 7, are ranked ninth and twelfth respectively. The phoric pronouns *it*, *something*, *nothing*, *anything,* and *whatever* are also found dispersed throughout the top fifty. Additionally, collocates confirming one of Partington's (2004, p. 140) primary findings that HAPPEN tends to be found in environments expressing "non-factuality", are found throughout this list: modal auxiliaries used in conditional expressions (*will, would, can, could,* and *might*); the conjunction *if*; and the question words *what* (ranked first by t-score), *when* and *why*.

Corpus data presented in this section illustrates that reliance on computer-generated Lists of collocates found within the 4:4 span is, at very least, potentially problematic. Again, the negative semantic prosody of CAUSE is readily apparent in Table 6.3 in statistically strong collocates like *problems*, *damage*, *cancer*, *death*, and so on. But no indications of positive/negative semantic prosody of HAPPEN are apparent in Table 6.4, created using exactly the same methods. The statistically strong lexical collocates of HAPPEN — *things*, *thing*, *know*, *think* — are not in themselves evaluative. If not for previous studies (Bublitz 1996; Partington 2004; Sinclair 2003) arguing otherwise, we might argue on this evidence alone, that HAPPEN

does not have a positive/negative semantic prosody.

6.2.3 BoE Pictures of CAUSE and HAPPEN

This section briefly discusses collocational evidence for semantic prosody as observed via the BoE Picture software. Table 6.5 shows the top twenty-five collocates in the 4:4 BoE T-Picture for CAUSE, and, unsurprisingly, evidence for the negative semantic prosody of CAUSE is again quite strong.

Table 6.5 Top 25 collocates in the BoE 4:4 T-Picture of CAUSE tagged as a verb, negative collocates bold

	N-4	N-3	N-2	N-1	N+1	N+2	N+3	N+4
1	the	the	which	can	by	to	to	of
2	that	that	that	has	A	**damage**	in	in
3	of	**damage**	it	which	the	**problems**	of	**problems**
4	**damage**	of	this	that	**problems**	**concern**	**problems**	**damage**
5	or	which	may	could	him	among	**damage**	to
6	which	this	likely	have	an	**harm**	**harm**	the
7	in	**problems**	can	to	some	stir	among	driving
8	**problems**	these	**virus**	is	them	lot	or	or
9	any	**pain**	could	would	**trouble**	**problem**	and	**pain**
10	**hiv**	may	has	may	serious	**pain**	**pain**	**trouble**
11	**disease**	because	**damage**	will	more	much	stir	fall
12	these	factors	they	what	**cancer**	**trouble**	fall	among
13	**pain**	**loss**	enough	had	**damage**	**death**	**dangerous**	**death**
14	such	**disease**	what	been	any	bodily	**lose**	rise
15	**desist**	or	does	was	**concern**	**disease**	**loss**	**crisis**
16	**loss**	**distress**	would	**damage**	me	**loss**	**trouble**	**disease**
17	this	**virus**	**problems**	also	her	**deaths**	drop	become
18	**distress**	thought	might	are	us	a	**disease**	**lose**
19	apologise	**blood**	have	be	**death**	sensation	when	**deaths**
20	because	**problem**	**disease**	not	**severe**	**upset**	change	blood
21	**problem**	**injury**	known	their	such	**storm**	**death**	**harm**
22	charged	**suffering**	going	it	**havoc**	great	**concern**	when
23	blood	**trouble**	any	might	considerable	**cancer**	than	**distress**
24	high	**hiv**	did	**problems**	**chaos**	by	**distress**	**collapse**
25	**lack**	believed	intent	**cancer**	**aids**	**injury**	but	**loss**

However, the BoE T-picture for HAPPEN in Table 6.6 below is, again, almost completely devoid of evaluative collocates. The table shows only the top twenty-five collocates[1], but it should be noted that there are only five unique negative collocates out of the 400 in the full picture (fifty

collocates in eight positions): *worst* (occurs twice, once ranked twenty-fourth at N-4, and once ranked forty-third at N-3), *terrible*, *accident*, *incident*, and *crash*, all ranked lower than twenty-fifth. In addition, there is one positive collocate, *best* at N-4.

Table 6.6: Top 25 collocates in the BoE-generated T-Picture (4:4 span) of HAPPEN, negative collocates bold, positive collocates underlined

	N-4	N-3	N-2	N-1	N+1	N+2	N+3	N+4
1	what	that	what	what	to	the	i	i
2	that	it	it	it	in	be	time	t
3	t	what	that	has	when	you	past	it
4	something	this	going	to	if	me	t	you
5	it	know	this	had	at	i	it	when
6	but	something	if	that	<p>	that	was	we
7	this	about	something	s	again	them	when	happen
8	if	if	things	t	I	him	we	is
9	know	things	see	will	but	we	ago	not
10	think	thing	when	this	here	us	you	happened
11	tell	t	nothing	have	next	my	next	that
12	thing	but	as	would	on	her	don	know
13	things	see	thing	whatever	now	it	people	was
14	don	is	doesn	things	before	this	last	ago
15	i	s	anything	is	because	quickly	if	think
16	nothing	when	about	never	so	your	did	few
17	not	believe	know	can	after	there	is	if
18	never	think	didn	just	there	our	future	there
19	best	how	could	could	with	said	there	time
20	find	just	did	was	during	know	country	don
21	wait	why	make	not	and	often	what	my
22	happened	knows	has	nothing	then	they	but	what
23	you	exactly	would	ever	we	all	life	s
24	**worst**	knew	won	something	it	night	just	just
25	why	nothing	exactly	might	is	mm	happened	but

As in the T-List shown in the previous analysis, the T-Picture contains a preponderance of phoric pronouns (*it*, *something*, *anything*, *whatever*, etc.), general nouns (notably *thing* and *things* are both found in all four columns from N-1 to N-4), question words (*who*, *what*, *where*, *when*, *why*, *how*), and modal auxiliary verbs. Once again, no argument for negative semantic prosody can be supported on this evidence alone, although both *thing* and *things* (shaded in Table 6.6), are central to phraseological studies that follow in Chapters 7 and 8, and in section 8.3.4.3 it is shown that *thing* and *things* can act as "evaluation carrier" (*cf*. Hunston and Francis,

2000, p. 134; Mahlberg, 2005, p. 152).

6.2.4 Grouping evaluative collocates by their total frequencies

This section looks more closely at frequency data in observations of the semantic prosody of CAUSE[2]. Barnbrook et al. (2013, p. 79) note that in addition to carefully selecting an appropriate statistical measure and span in collocational studies, "[…] there is the question of the threshold value, which can usefully filter out rare words which would otherwise dominate the output, the 'long tail' of the Zipfian distribution of words[3]." In other words, they suggest that low-frequency collocates can be excluded from consideration. However, their comments are not made in reference to the collocational nature of semantic prosody, so it is unclear what status the long tail of collocates might have in observations of evaluation. As shown in Chapter 2, Stubbs (1995, p. 14) is also aware of the apparent necessity of acknowledging (admittedly arbitrary) frequency thresholds, and suggests that in cases where a node has a large number of low-frequency, statistically insignificant evaluative collocates "[i]t may […] be worth grouping the data." It is argued in this section that grouping evaluative collocates by summing their FaC values does indeed seem to be a helpful approach. Calculation of a combined t-score, as Stubbs suggests, may not be necessary, however.

The argument for grouping collocates can be illustrated by looking at profiles created using different statistical methods. Table 6.7 shows the top-twenty N+1 collocates from the Pictures of CAUSE in the BoE ranked by raw frequency, T-score, and MI score. It was discussed in Chapter 2 that lists of collocates ranked by MI score tend to favour very infrequent words, and the researcher's intuition alone is likely sufficient to judge that the FaCs in the MI list are too low to support a convincing argument for semantic prosody. The MI data is presented here to allow for a comparison of extreme FaC values to illustrate an argument that will be central to

analyses that follow.

First, of the collocates ranked by MI, sixteen are negative, and context reveals that the remaining four are found in predominantly negative propositions (i.e. *masculinisation*, *photosensitivity*, and *drowsiness* in this context are very often unwanted side-effects of certain medications, and *untold* is most frequently followed by *damage*, *misery*, *suffering*, *harm*, *stress*, etc.). In comparison, there are only four negative collocates in the Frequency List, and seven in the T-list.

Table 6.7 Comparison of the top-twenty collocates of CAUSE (89,830) at N+1 in the BoE Pictures ranked by RAW frequency, t-score, and MI score

	Ranked by Frequency		Ranked by T-score			Ranked by MI Score		
	Collocate	FaC	Collocate	T-Score	FaC	Collocate	MI	FaC
1	by	15,283	by	120.08	15,283	masculinisation	10.28	3
2	the	9,004	a	66.14	7,848	**grievous**	10.22	195
3	a	7,848	the	42.58	9,004	**grevious**	10.17	6
4	**problems**	1,488	**problems**	38.01	1,488	**consternation**	9.81	148
5	an	1,471	him	31.97	1,250	**ructions**	9.80	24
6	him	1,250	an	30.73	1,471	photosensitivity	9.76	4
7	it	1,183	some	30.41	1,169	**havoc**	9.61	362
8	some	1,169	them	29.46	1,114	**onchocerciasis**	9.53	4
9	them	1,114	**trouble**	25.07	646	**irreparable**	9.47	56
10	to	926	serious	24.56	628	drowsiness	9.19	33
11	more	899	more	23.47	899	**uproar**	9.11	149
12	you	773	**cancer**	23.15	551	**apoplexy**	9.07	16
13	her	755	**damage**	22.41	516	untold	8.87	74
14	any	649	any	21.96	649	**chagas**	8.86	4
15	**trouble**	646	**concern**	20.83	449	**elephantiasis**	8.51	3
16	**serious**	628	me	20.62	614	**anaphylactic**	8.41	3
17	me	614	her	20.45	755	**malformation**	8.30	4
18	us	567	us	20.32	567	**mayhem**	8.27	124
19	cancer	551	**death**	19.86	432	**ulceration**	8.27	6
20	**damage**	516	**severe**	19.50	387	**dioxins**	8.26	16

However, Table 6.8 below shows that the number of evaluative collocates is not necessarily the most salient detail in observations of semantic prosody. Examination of the FaC values reveals a situation in which a substantial number of negative collocates in the MI list accounts for only a very small number of the total occurrences of the node. Again, this is unsurprising in itself,

but it is illustrative of the argument that summed FaC values presented as percentages of total frequencies can be revealing measures of significance of groups of collocates.

Table 6.8 shows the four negative collocates at N+1 of the Frequency Picture have a combined FaC of 3,201, which is 3.6% of the total number of 89,830 occurrences of CAUSE tagged as a verb in the BoE. The seven negative collocates in the t-score list account for 4,469 lines, or 5.0% of all occurrences. In stark contrast, the sixteen negative collocates by MI account for 1,120 lines, or just 1.2% of the occurrences of CAUSE in the BoE. Another way of describing this incongruity is that there are more than twice as many negative collocates by MI as there are by t-score, but the t-score collocates occur four times as frequently in the corpus.

Table 6.8 Total FaC and total negative collocate FaC of the top-twenty collocates of CAUSE (89,830) at N+1 ranked by frequency, t-score, and MI score

	Freq	T-score	MI Score
Total FaC	47,334	45,720	1,234
FaC % Of total occurrences	52.7%	50.9%	1.4%
Number of neg collocates	4	7	16
Negative collocates total FaC	3,201	4,469	1,120
Negative FaC % of total FaC (47,334)	6.8%	9.8%	90.8%
Total Negative FaC as % of 89,830 Occurrences	3.6%	5.0%	1.2%

In theory, of course, the converse is also possible; a small number of collocates could account for a large number of occurrences (see Chapter 2 for an example). How these discrepancies are interpreted could conceivably have significant effects on arguments for or against claims that an item has a specific semantic prosody, and throughout the chapters that follow, evaluative FaC totals will be central to analysis and conclusions.

6.3 The Collocate's Relationship to the Node

In this section, it is argued that defining a collocate's syntactic and semantic relationship to the

node are central to observations of semantic prosody. The majority of evaluative collocates observed in the smaller, standard, 4:4 window are likely to be found in strong syntactic and semantic relationships with the node. Noun collocates in close proximity to the verb CAUSE, for example, are likely to be its grammatical subjects and objects. The further afield collocates are recovered, however, the greater the necessity of justifying their syntactic and semantic relationships with the node.

6.3.1 Focussing the investigation

Thus far, the methods of observing collocational evidence of the negative semantic prosody of HAPPEN have been almost entirely unsuccessful. However, since the focus is on the grammatical subjects of the verb HAPPEN (we are essentially looking to confirm or deny that *what happens* is negative), the corpus can be interrogated with queries designed to return only nouns in relevant positions within the span. Four such queries[4] were used, and the results combined into a positional frequency table (PFT), as shown in Table 6.9.

It is possible to make at least four important observations regarding the collocates in Table 6.9. First, both *thing* and *things* (highlighted in the table) stand out as the most frequent collocates in each of the four positions shown. These collocates are central to phraseological analyses in Chapter 7.

Secondly, Table 6.9 does initially appear to support the claim that HAPPEN has a negative semantic prosody; twenty-eight of the top fifty nouns at N-1, for example, are negative, and only two are positive. However, there is a sizable drop in the number of negative collocates to only thirteen at N-2, and this tendency toward fewer and fewer negative collocates continues: there are only six at N-3 and five at N-4. This extreme drop suggests that the

Table 6.9 Positional T-Score table of relevant columns for four corpus queries isolating noun collocates from N-1 to N-4 of HAPPEN in the BoE

	N-4	N-3	N-2	N-1	FaC
1	thing	things	things	things	2,177
2	things	thing	thing	thing	1,112
3	time	time	doesn	**accident**	603
4	people	idea	time	**incident**	386
5	way	sort	matter	**crash**	122
6	sort	way	way	**attack**	122
7	years	doesn	incident	**tragedy**	112
8	fact	accident	accidents	**accidents**	109
9	kind	years	events	events	105
10	lot	chance	lot	event	64
11	course	day	look	**explosion**	61
12	question	disaster	accident	change	63
13	world	erm	case	miracle	57
14	idea	story	event	**shooting**	51
15	doesn	events	course	changes	52
16	life	kind	changes	**disaster**	45
17	day	question	day	stuff	45
18	erm	life	opposite	drama	44
19	terms	course	miracles	**incidents**	42
20	god	lot	earth	fact	47
21	case	fact	change	**hell**	39
22	place	chances	fact	**unthinkable**	35
23	sense	people	erm	opposite	31
24	story	account	stuff	**shit**	29
25	picture	changes	hell	**attacks**	30
26	year	world	question	marketing	27
27	view	incident	disaster	**horror**	25
28	chance	truth	tragedy	process	27
29	night	night	idea	life	36
30	war	event	life	revolution	24
31	fear	stranger	q	miracles	23
32	truth	example	moment	**collision**	22
33	days	responsibility	incidents	action	26
34	police	war	attack	**abuse**	22
35	responsibility	matter	war	**injury**	22
36	home	case	process	**violence**	20
37	example	place	miracle	**mistakes**	19
38	events	days	sort	**blast**	18
39	questions	view	story	earthquake	17
40	incident	god	night	reverse	18
41	situation	year	unthinkable	**murder**	16
42	accident	accidents	chance	**murders**	19
43	game	terms	crash	experience	22
44	event	change	reverse	**war**	16
45	moment	accounts	holocaust	earth	14
46	look	tragedy	deal	**smash**	16
47	attention	game	mistakes	beat	18
48	order	months	people	name	18
49	reason	details	world	**injuries**	14
50	horror	reason	revolution	**death**	17
				total FaC	6,099
				total neg FaC	2,076

negative semantic prosody is not nearly as strong for HAPPEN as it is for CAUSE (which shows no such decline in numbers of negative collocates). It could be that the negativity of HAPPEN is not in fact activated primarily via its grammatical subjects; a small study focussing on other parts of speech (specifically adjectives modifying the subjects, and adverbs modifying HAPPEN itself), however, did not reveal convincing evidence for semantic prosody. The likeliest hypothesis, mentioned previously (Section 6.2.2), is argued by Bublitz (1996, p. 14), who suggests that the grammatical subjects of HAPPEN are either structurally complex or are phoric words, the referents of which are found beyond the standard span.

A third important observation is that, as contended in 6.2.4, it is difficult to sustain an argument that the total number of evaluative collocates is a strong indicator of an item's semantic prosody when the summed negative FaC values presented as a percentage of the total number of occurrences of HAPPEN in the corpus is low. Table 6.10 appears to confirm that the data in Table 6.9 may not be as strong as it at first appears.

Table 6.10 Comparison of FaC raw totals and percentages for the top fifty collocates and the negative noun collocates at N-1 for the BoE query for noun+HAPPEN

		% of Total	% of noun+ HAPPEN
Total occurrences of HAPPEN in the BoE	149,408	—	—
Lines returned for HAPPEN preceded by a noun	10,554	7.1%	—
Top-fifty FaC total	6,099	4.1%	57.8%
Negative collocates in top-fifty FaC total	2,076	1.4%	19.7%

Table 6.10 shows that the query for HAPPEN immediately preceded by a noun returns 10,554 lines. This is only 7.1% of the 149,408 instances of HAPPEN in the BoE. The total FaC of the top fifty noun collocates is 6,099, which is 57.8% of the focussing query concordance, but only 4.1% of HAPPEN in the BoE. Most importantly, the table shows that the total negative noun FaC

at N-1 is 2,076. This means that these twenty-eight negative collocates occur at N-1 in only 1.4% of the occurrences of HAPPEN in the BoE.

For comparison, Table 6.11 shows the results of similar frequency analysis of collocates at N+1 for lines containing CAUSE immediately followed by a noun in the BoE. The table show that the BoE contains 22,433 instances of CAUSE followed by a noun.

Table 6.11 Comparison of FaC raw totals and percentages for the top fifty collocates and the negative noun collocates at N-1 for the BoE query for CAUSE+noun

		% of Total	% of noun+ CAUSE
Total occurrences of CAUSE in the BoE	89,830	—	—
Lines returned for CAUSE followed by a noun	22,433	25.0%	—
Top-fifty FaC total	9,871	11.0%	44.0%
Negative collocates in top-fifty FaC total	9,010	10.0%	40.2%

The top fifty collocates at N+1 account for 9,871 (44.0%) of these instances, or 11.0% of the total occurrences of CAUSE. Of the top fifty, forty-two are negative and these have a combined FaC of 9,010, which is 40.2% of the 22,433 occurrences returned by this focussing query, and 10.0% of the total occurrences of CAUSE. This is, of course, considerably higher than the 1.7% negative FaC of nouns preceding HAPPEN.

Finally, the fourth observation (as mentioned in the discussion above of Table 6.3 which shows the T-List of collocates of CAUSE) is that some of the ostensibly neutral collocates are in fact often found in phrasal structures, or "structurally complex collocates" (Bublitz, 1996, p. 14), that evaluate negatively. For example, 417 of the 458 instances of *bodily* as a collocate of CAUSE, ranked 41st in the T-List, are part of the phrase *bodily harm*, and 204 of these are *grievous bodily harm*. For another of these apparently neutral collocates, *blood*, we see not only individual negative collocates (*cause blood clot*(*s*)/*poisoning*/*feuds*/*abnormalities* etc.), but longer

evaluative phrases start to emerge like *blood pressure problems*, and *blood vessel spasms*. Bublitz (1996, p. 14) argues that "the majority" of collocates of HAPPEN are structurally complex, which could serve to explain why there is virtually no evidence of semantic prosody in the T-List or T-Picture, and why the collocates in the focussing query PFT have such a low total FaC.

However, focussing queries like those used in the construction of Table 6.9 tend to overlook these complex collocates because, of course, they include only one part of speech, in this case nouns. Complex collocates necessarily comprise multiple mixed elements from relatively simple foundations of modification in the form of adjectives and adverbs to full clause structures. What the table does suggest, though, is that the node itself can become a part of a larger, complex phrasal structure. For example, Table 6.9 shows *accident* and *disaster* at N-3. Not shown is their collocate FaCs of 121 and 98 respectively. The lexical phrase *an accident waiting to happen* is found in the BoE sixty-six times, which accounts for more than half of the 121 instances of *accident* in this position. Similarly, *a disaster waiting to happen* is found sixty times, or more than 60% of the occurrences of *disaster* at N-3. The implication is that these are not in fact collocates, but rather elements of the phraseological behaviour of HAPPEN. Chapter 8 discusses these two phrases and the implications of this frequency data in much more detail.

6.3.2 Syntactic and textual relationships

This section discusses how an evaluative collocate's relationship to the node evinces its evaluative mode. In Section 2.6, Mahlberg's (2005, p. 149) notion of the evaluative cline of "increasing context dependency" was discussed in detail. There, it was shown that semantic prosody is one mode of evaluation that can be placed in the middle of a cline of evaluative meaning. On one end of the cline is core evaluative meaning, which requires no additional

contextual or linguistic knowledge to be observed. On the other end is textual evaluation, which not only usually requires the most linguistic context, it often requires "the conceptual approach" to evaluation, and can be observed, at times, even in the absence of overtly evaluative language (Mahlberg, 2005).

That these types of meaning exist on a continuum of contextual dependency means it is impossible to mark a clear cut-off point, i.e. a specific span, where evaluation ceases to be prosodic and becomes textual. A strict adherence to the 4:4 span is not necessary since strong syntactic relationships can be expressed over very short spans (i.e. adjective-noun pairings) or over very large spans (as is sometimes the case with pronouns and their antecedents).

Some examples of various kinds of strong syntactic relationships are shown in the lines that follow. In these four lines taken from the BoE, the primary evaluative collocate is the grammatical subject of the node in line 1; the direct object in 2; the indirect object(s) in 3; and both subject and object in 4. Other syntactic relationships are, of course, possible.

1. has only just begun. Stress **causes** complex changes in blood
2. The mystery of how power lines could **cause** cancer is, it seems, as far
3. of a sort that would probably **cause** us discomfort or embarrassment
4. Psychological disorders **cause** emotional distress,

However, evaluation is considered textual where relevant evaluative words or phrases show either no clear syntactic relation to the node, or where the amount of context becomes too large to sustain a convincing argument for collocation.

5. Its basic premise is that psychological problems arise when people try to interpret (a cognitive activity) what **happens** in the world on the basis of irrational beliefs.

In line 5, *problems* is found at N-11 which is well beyond the standard span. In itself, this does

not necessarily disqualify it as evidence of semantic prosody. But in this instance, the syntactic relationship has reached a level of complexity that no longer evinces prosodic evaluation. That is, negative evaluation is present but is considered textual.

6.4 Qualitative analyses of 500-line concordances for CAUSE and HAPPEN

This section discusses the qualitative analysis of 500-line concordances of CAUSE and HAPPEN in the BoE (see Section 5.5 for detailed discussion on the differences between quantitative and qualitative analyses in this thesis). It is demonstrated in this section that qualitative analysis is required to establish whether the relationship between collocate and node is relevant to claims of semantic prosody. Furthermore, other important syntactic and semantic factors, for example structurally complex collocates and contexts expressing idiomatic meanings, can usually be identified only qualitatively.

6.4.1 Qualitative analysis of 500 lines of CAUSE in the BoE

Qualitative analysis of the CAUSE concordance identified lines containing single word collocates far beyond the 4:4 span but with a strong syntactic relationship to the node, as in line 6, where *illness* is not syntactically related to *cause*, but *trouble* is the indirect object:

6. The illness persisted, and Aunt O-hana's demands continued to **cause** everyone in the family a great deal of trouble,

Also revealed are evaluative phrasal collocates comprising two or more neutral words, as in the following lines.

7. physics by which greenhouse gases **cause** warming is uncontested
8. This greenhouse effect is predicted to **cause** an overall global warming
9. of the throat and stomach, thereby **cause** side effects.
10. produced under the skin builds up and **causes** the milk to boil over.

Some of these phrases might at first appear to be evidence of textual evaluation in that they seem to require additional linguistic or contextual knowledge for their evaluative polarity to be observed. As Mahlberg (2005, p. 150) argues, however, "[…] a text meaning may develop into a prosody." The phrase *greenhouse gases* is a likely candidate for such a transformation, namely a meaning that began as textually negative because it required extra-linguistic knowledge of what these gases were and how they were affecting the environment has now become almost universally well-understood and considered negative. For this reason, this and similar phrases are considered in the current study to evince semantic prosody, not textual evaluation.

Qualitative analysis also allows for disambiguation among senses of collocates. The word list used to automatically identify negative collocates Section 5.5.1 does not contain *friction*, for example. In the 448-line concordance (fifty-two mis-tagged lines were removed from the 500-line sample), however, CAUSE *friction* is found three times, and in each instance varying degrees of context reveal that the *friction* is clearly negative:

11. when the children were young it **caused** friction. Douglas was often cruelly tormented by jealous siblings.
12. and a special envoy from Pakistan, Mr Shareyar Khan, over issues **causing** friction between the two countries. These include Kashmir and the arms race.
13. Maybe even to her detriment. Long-time rival Corbett described Gilbert as aloof and claimed the other competitors wanted her to beat Gilbert in the first Meadow Lea series. It **caused** friction. Friction which is still evident. When asked if there was any angst between her and former Meadow Lea number one Corbett, Gilbert paused before answering: `Not really" – although their less

The final results of the qualitative analysis are shown in Table 6.12. The table shows that a total of 381 (85.0%) lines were deemed to evaluate via negative semantic prosody. This is sixty lines more than the 321 that were shown in Section 6.2.1 to have a negative collocate in the 4:4 span. Only twelve lines (2.7%) were found to evaluate positively, and fifty-four are neutral/unknown.

Table 6.12 Results of qualitative analysis of 448 random lines of CAUSE tagged as a verb in the BoE

	Frequency	%
Negative Semantic Prosody	381	85.0%
Positive Semantic Prosody	12	2.7%
Negative Textual Evaluation	1	0.2%
Positive Textual Evaluation	0	0.0%
Neutral/Unknown	54	12.1%
Total	448	100%

Table 6.12 also shows that only one line was judged to evince negative textual evaluation. In Line 14, the collocate *shank* is considered neutral in itself, and only the specialist knowledge provided by *clubhead* and the fact that it appears to be something we would want to *cure* indicates negative evaluation[5], and so here it is considered textual.

14. and the clubhead is thrown onto an outside path. This action also **causes** a shank. Let's see if we can cure, or at least reduce the slice.

6.4.2 Qualitative analysis of 500 lines of HAPPEN in the BoE

In anticipation of the difficulties encountered by Bublitz (1996) in identifying collocates of HAPPEN (see below for detailed discussion), the analysis discussed in this section used *ca.* 200 character lines rather than the 10:10 span employed earlier. Following Bublitz (1996, pp. 17–18), Partington (2004, p. 136), and Sinclair (2003, p. 125) thirty-six lines expressing the "by chance" meaning of HAPPEN were initially removed and analysed separately. However, these lines were found to evaluate similarly to the rest of the instances in the concordance, so, in the end, they are included in the results of this analysis.

Table 6.13 shows that only 20.7% of the lines have negative collocates in close syntactic and semantic relationships to the node and are thus considered evidence of semantic prosody. Further, as Table 6.13 shows, 30.4% of the lines contain evidence of negative textual evaluation. Also noteworthy, however, is that 34.7% remain neutral/unknown.

Table 6.13 Results of qualitative analysis of the 464-line BoE Concordance of HAPPEN

Evaluation Type and Polarity	Frequency	% of Total
Negaive Semantic Prosody	96	20.7%
Positive Semantic Prosody	26	5.6%
Neutral/Unknown	161	34.7%
Negative Textual Evaluation	141	30.4%
Positive Textual Evaluation	40	8.6%
Total	464	100.0%

Because the data in Table 6.13 makes clear distinctions between prosody and textual evaluation, it is worth looking briefly at some examples of each from the concordance. To begin, evidence of negative semantic prosody is observable in lines 15 and 16 in the grammatical subjects of HAPPEN found within the 4:4 span:

15. `We can't help thinking something sinister has **happened**.
16. `A terrible thing **happened**, Ann. Your close friend was killed in an accident.

Line 17 shows how adverbial modification can also activate the negative prosody of HAPPEN:

17. As <p> we've seen, it could **happen** violently and anti-democratically and that could be catastrophic.

In 17, neither the subject of *happen* (*it*) nor its antecedent (*change*) are intrinsically evaluative, but *violently* and *anti-democratically* constitute strong collocational evidence of negative semantic prosody.

One of the difficulties with a collocational definition of semantic prosody that requires a clear syntactic relationship between collocate and node (even regardless of span) is observed when a node frequently has pronouns and general nouns as grammatical subjects. As Bublitz has noted, the grammatical subject of HAPPEN is very often a pronoun, the antecedent of which is found well beyond the standard span of investigation, or a structurally complex item. It is worth

quoting Bublitz at length to underscore the importance of this difficulty:

> While it is easy to draw up a long list of single nouns as collocates of *cause* from the data […] (*inflation*, *pain*, *disappointment*, *trouble*, *bombardment* etc), this is hardly possible for *happen*. Of the 303 occurrences, there are only very few single word collocates which display a positive or negative semantic prosody: *delay*, *problem*, *thing*, *accident*, *encroachments*. Some are two word collocates (*shocking thing*, *something dreadful*, *something gastly* [sic], *nothing odd*), but the majority consists of structurally complex collocates or else of (usually anaphoric) pronouns. To learn the reference of the latter, we sometimes have to go well beyond a 4:4 or even 8:8 span.

Line 18, an example of the problem described by Bublitz, has been labelled 'negative semantic prosody' because *violence*, despite occurring at N-14 (far beyond the standard span) is the obvious antecedent of *that*, the grammatical subject of *happen*. Similarly, the antecedent of *that* in 19 appears to be *he might be recaptured*, a phrase evincing the negative prosody of HAPPEN.

18. out of the game, but the violence we saw is unacceptable, and we have to make sure that does not **happen** again."
19. he might be recaptured, and he'd waited too long for freedom to let that **happen**.

Many lines in this concordance, however, display what is here called textual evaluation, as in the following examples:

20. to allow the dreaded truth into his consciousness, he had forgotten nearly everything. It had **happened**, after all, thirty-six years before.
21. it means deportation internment, and outright murder. I should not have been surprised by what **happened** in Koreatown or by the ignorance and hatred
22. feared that I might telephone Alan Walters, who was in America and quite oblivious to what was **happening**, and that Alan would resign. This would have deprived him of the excuse he wanted.

Line 20 is an example of 'negative textual evaluation' because the phrase *the dreaded truth*, at N-13 to N-11, allows us to infer the negative polarity of this stretch of text as a whole, but what exactly *happened* — i.e. the antecedent of the grammatical subject *it* — is not revealed even in the *ca.* 200 characters examined. Likewise, despite the presence of many negative words in 21

(*deportation*, *internment*, *murder*, *ignorance*, *hatred*), none of these expresses *what happened in Koreatown*. The same is true in 22 where neither *feared*, *oblivious*, nor *resign* are in a close enough relationship to *what was happening* to call them evidence of semantic prosody.

6.5 Conclusion

This chapter has demonstrated that claims that semantic prosody is a collocational phenomenon necessarily beg the question of what is meant by the term "collocation" itself. Although evidence for semantic prosody is, in fact, often observed in groups of statistically strong collocates, corpus evidence shows that statistical significance is not a requirement of such evidence. Corpus data show that semantic prosody is "collocational" only in the broadest, simple co-occurrence sense. In addition, the results presented and discussed in this chapter demonstrate that the term "collocate" in the context of corpus observations of semantic prosody must include the notion of close syntactic relationship and allow for structural complexity.

Section 6.2 illustrates that different statistical measures highlight different sets of collocates. McEnery, Xiao, and Tono (2006, p. 215) argue "[r]aw frequency is a poor guide to collocation", but this does not necessarily entail that raw frequency is an equally poor guide to semantic prosody. In fact, corpus data suggest that the most straightforward of the statistical measures, Frequency as Collocate (FaC) very often reveals striking differences between individual collocates and collocate sets. Specifically, corpus data suggests that the number of evaluative collocates is not necessarily as important to observations of semantic prosody as the total frequencies of the sets of evaluative collocates. This is especially important in analysis that follows, where, for example, items are examined with similar numbers of positive and negative collocates, but in every case, one of the sets has a substantially higher total FaC.

Section 6.3 focussed on the collocate's relationship to the node. In that section, corpus data demonstrated that the semantic prosody of an item is activated only when the collocate in question is in close syntactic association with the item (i.e. as the grammatical subject or object of a verb, as modifier within a noun or verb group, the antecedent of a pronoun, etc.). The fact that semantic prosody is found on an evaluative cline between "core evaluative meaning" and "textual evaluation" was demonstrated by showing that in many cases, HAPPEN is found in evaluatively negative environments, but that this environment is textual in the sense that it requires additional linguistic or extra-linguistic knowledge to be activated.

Finally, in Section 6.4 results of qualitative examinations of concordance of CAUSE and HAPPEN were presented. Results showed that, unsurprisingly, CAUSE appears to evaluate via semantic prosody in more than 80% of the lines. In contrast, concordance analysis showed that propositions involving HAPPEN appear to evaluate via negative semantic prosody only *ca.* 20% of the time. A further *ca.* 30% were observed to evaluate via negative textual evaluation. Almost 35% remained neutral/unknown.

An unforeseen result of examining various methods of revealing evaluative collocates of CAUSE and HAPPEN is that the semantic prosody of HAPPEN noted by Bublitz (1996), Partington (2004), and Sinclair (2003) and cited by many publications on semantic prosody (Louw, 1993, p. 158; Stubbs, 1995, p. 3; Cotterill, 2001, p. 292; Tognini-Bonelli, 2001, p. 111; Whitsitt, 2005, p. 287; Xiao and McEnery, 2006, p. 106; Bednarek, 2008, p. 121; Granger and Paquot, 2008, p. 31; Kennedy, 2008, p. 36; Ellis, Frey and Jalkanen, 2009, p. 90; Stewart, 2010), appears to be, at best, much weaker than that of CAUSE, and at worst, not observable in the data at all. At very least, the data supports Bublitz's (1996, p. 19) conclusion: "[S]emantic prosody is more stable or less stable, in other words, the strength of association between node and collocate based on

semantic prosody is a matter of degree." Partington (2004, p. 153) arrives at a similar conclusion: "Some items have a stronger good or bad prosody than others." The semantic prosody of CAUSE is observable regardless of how the collocational profile is constructed, but the semantic prosody of HAPPEN is hardly apparent using identical methods. Instead, specialized focussing queries which return only a small fraction of the total instances, are required to observe what amounts to very little evidence of the negative prosody of HAPPEN in a collocational profile. If semantic prosody is collocational, and if we define 'collocate' as a single word that co-occurs with the node more often than random chance would suggest, then, quite simply, HAPPEN does not appear to have a negative semantic prosody. Even if we broaden the definition of collocate to include much larger spans and relax the notion of statistical re-occurrence, evidence that HAPPEN has a negative semantic prosody is very sparse.

Notes

[1] Showing only twenty-five collocates is a somewhat arbitrary cut-off point chosen because it is thought to display sufficient evidence to support the argument while not overwhelming the page with huge lists of words.

[2] Comparative analysis to HAPPEN is not illuminating because, as we have seen, there are not enough evaluative collocates in the BoE Lists and Pictures.

[3] Zipf's Law states that in large corpus the frequency of a given word is inversely proportionate to its rank in the frequency list. A word ranked N on the list will have a frequency of approximately 1/N. This means that the word ranked second in the list will occur approximately half as frequently as the most frequent word; the third most frequent word will occur about one-third as frequently, and so on.

[4] NOUN+happen@; NOUN+1,1happen@; NOUN+2,2happen@; NOUN+3,3happen@

[5] The negative, golf-related meaning, of *shank* is the fifteenth sense at https://www.collinsdictionary.com/dictionary/english/shank (accessed august 16, 2017), and is not found at all in the third edition of CCED (2001, p. 1425), which contains only two senses.

CHAPTER 7:
EFFECTS OF PHRASEOLOGICAL BEHAVIOUR ON THE SEMANTIC PROSODY OF HAPPEN

This chapter examines corpus evidence of the evaluative modes of three phrases comprising the wordform *happen* and the general nouns *thing* and *things*, specifically the "nested" collocation (*cf*. Hoey 2005) *things happen* and the lexical phrases *these things happen* and *make things happen*. Corpus data show that the evaluative environment that HAPPEN is found in is greatly affected by its phraseological behaviour.

First, analysis of the short phrase *things happen* in Section 7.1 demonstrates that even ostensibly very small phraseological changes, i.e. from *HAPPEN* to *things happen*, appear to have substantial effects not only on an item's dominant evaluative polarity, but also the mode — core, prosodic, or textual — by which it evaluates. The chapter then shows that phraseologies sharing this common core *things happen*, namely *these things happen* (Section 7.2) and *make things happen* (Section 7.3) evaluate differently.

Each section of this chapter includes discussions of collocational profiles that reveal problematic tensions between numbers of evaluative collocates and their total Frequency as Collocate values (FaC); the effect of requiring collocational significance (in this case t-scores above 2.0); the problematic nature of a profile containing many collocates that occur only once each; and the difficulty in ascertaining the logical/syntactic relationship between the collocates in the profile and the node. Each section also includes discussion of a qualitative analysis of a sample concordance for each phrase.

7.1 Things Happen

The short phrase *things happen* (1,497) was selected for examination because of the apparent

salience of the collocate *things*, indicated by its high frequencies in collocational profiles of HAPPEN and its word forms. Table 6.4 above shows that *thing* and *things* are ranked ninth and twelfth respectively in the BoE T-List for HAPPEN, and are conspicuously the only lexical noun collocates in the list; they are also ranked highly in each of the 4:0 positions of the T-Picture; and in Table 6.9, which isolates noun collocates at each position in the 4:0 span, the top two collocates in each position are *thing* and *things*.

7.1.1 Collocational profile of *things happen*

This section discusses the evaluative mode and polarity of the nested collocation *things happen*. Corpus data illustrate the importance of comparing combined evaluative FaC values before proclaiming that an item has a positive or negative semantic prosody based on the number of evaluative collocates in a profile. The investigation focusses primarily on adjectives modifying *things* because these are the syntactically and logically relevant collocates expressing what kinds of *things happen*.

There are 626 lines[1] in the BoE in which a word tagged as an adjective immediately precedes *things happen*. Table 7.1 shows the top fifty of these adjective collocates ranked by FaC. The table also shows the collocates' t-scores. Negative collocates have been highlighted in bold, and positive collocates are underlined. Collocates with t-scores lower than the standard significance cut-off of 2.0 are shaded grey. The collocates in Table 7.1 account for 546 (87.2%) of the 626 occurrences in the concordance. This indicates that the data is representative of the overall trends and collocates ranked fifty-one and lower are not likely to skew the results greatly.

Table 7.1 shows that the negative to positive ratio of adjectives immediately preceding *things HAPPEN* is 14:15. It is worth noting that if the standard t-score cut-off of 2.0 were implemented,

eighteen of these top fifty collocates (shaded in Table 7.1) would not be considered. This would have the effect of lowering the negative to positive ratio to 8:10 (six fewer negative and five fewer positive). For this profile analysis, such a reduction is not especially problematic in that the balance of negative to positive is not substantially altered. However, as demonstrated later in this chapter, other profiles display very prominent differences if t-score thresholds are applied.

Table 7.1 Adjectives at N-1 of *things HAPPEN* in the BoE, ordered by frequency

	Collocate	FaC	T-score		Collocate	FaC	T-score
1	**bad**	69	8.23	26	various	6	2.42
2	strange	50	7.06	27	extraordinary	6	2.43
3	such	44	6.55	28	same	5	2.09
4	good	38	6.08	29	new	5	1.78
5	other	30	5.41	30	right	5	2.01
6	**terrible**	28	5.28	31	unexpected	5	2.23
7	**worse**	23	4.78	32	big	4	1.86
8	certain	19	4.33	33	**unfortunate**	4	2.00
9	interesting	18	4.23	34	dramatic	4	1.98
10	funny	12	3.45	35	**unpleasant**	4	2.00
11	**horrible**	12	3.46	36	nice	3	1.69
12	several	11	3.26	37	**worst**	3	1.70
13	different	10	3.09	38	special	3	1.65
14	odd	10	3.15	39	significant	3	1.70
15	exciting	9	2.99	40	unusual	3	1.71
16	**awful**	9	2.99	41	**ugly**	3	1.72
17	**dreadful**	9	2.99	42	**negative**	3	1.71
18	similar	8	2.79	43	magical	2	1.40
19	little	8	2.70	44	unbelievable	2	1.41
20	wonderful	8	2.81	45	marvellous	2	1.40
21	weird	8	2.82	46	amazing	2	1.39
22	**nasty**	8	2.82	47	**shocking**	2	1.40
23	important	7	2.57	48	startling	2	1.41
24	positive	7	2.62	49	surprising	2	1.40
25	great	6	2.32	50	**irrational**	2	1.41

This initial observation that *things happen* has an almost equal number of negative and positive adjective collocates at N-1 suggests a 'balanced' prosody. That is, *things happen* appears to be imbued equally with both positivity and negativity. Table 7.2, however, arranges thirty-nine of the adjectives from Table 7.1 by very general semantic preferences and FaC. This allows us to see more clearly the effects of summed FaC values. In addition to the 'generally negative' and

'generally positive' preferences, Table 7.2 shows a third preference for adjectives meaning 'strange or unexpected'. These collocates arguably do not have core evaluative meanings but would likely be found to contribute to positive-negative evaluation in larger contexts.

Table 7.2 Semantic preferences of the adjectives returned from the query ADJECTIVE *things* HAPPEN (626) in the BoE

Generally Negative		Generally Positive		Strange or unexpected	
Collocate	FaC	Collocate	FaC	Collocate	FaC
bad	69	good	38	strange	50
terrible	28	interesting	18	odd	10
worse	23	funny	12	weird	8
horrible	12	exciting	9	unexpected	5
awful	9	wonderful	8	dramatic	4
dreadful	9	important	7	unusual	3
nasty	8	positive	7	startling	2
unfortunate	4	great	6	surprising	2
unpleasant	4	extraordinary	6	unbelievable	2
worst	3	right	5		
ugly	3	nice	3		
negative	3	special	3		
shocking	2	magical	2		
irrational	2	marvellous	2		
		amazing	2		
Total FaC	179		128		86

However, the number of evaluative collocates is not necessarily indicative of the item's semantic prosody. As Table 7.2 illustrates, summing the FaC values of the negative and positive collocates reveals a ratio of 179:128 (or almost 3:2), meaning that these fourteen negative collocates account for almost 30% more instances than their fifteen positive counterparts. This suggests that, when characterizing what types of *things happen*, a language user's initial pragmatic choice, i.e. semantic prosody, is in fact much more often 'bad' than 'good' despite the almost equal number of negative and positive collocates.

However, it is often the case that a word is frequent in a corpus because it is an element of a frequent phrase (see for example, Stubbs 2007:164). The same can often be claimed of a node's high-frequency collocates, in that a specific collocate is often very frequent because it combines

with the node to form a frequently occurring phrase. For instance, when the initial pragmatic choice is to express a negative proposition, the collocate *bad* is chosen 38.5% of the time (69 out of the total bad FaC of 179); next is *terrible*, which is much less frequent at 15.6%. Likewise, when the initial pragmatic choice is to express a positive proposition, the collocate *good* is selected 29.7% of the time (38 out of the total good FaC of 128), followed by *interesting* at only 14.1%. This is potentially important because it could be that *bad* and *good* are much more frequent for reasons that are not observable in the collocational profile, namely that they are elements of frequent phraseologies (see Chapter 8 for further discussion of this notion and its implications).

Using the methods outlined in Section 2.4.4 we can calculate the expected probability of the phrase *things happen* (1,497) in the 450-million word BoE:

$$\frac{1{,}497}{450{,}000{,}000} = 0.000003$$

Similarly, we can calculate the probability of *bad* occurring in the corpus:

$$\frac{82{,}575}{450{,}000{,}000} = 0.0002$$

And now it is a simple matter of calculating the likelihood of *bad things happen*:

$$0.000003 * 0.0002 = 0.0000000006$$

This means we would expect to see *bad things happen* ca. 0.3 times in the 450-million word BoE. As we have seen, however, there are sixty-nine instances of the phrase in the corpus, just over 250 times more than the expected value.

By comparison, the word form *good* occurs 369,849 times in the BoE (a probability of 0.0008; four times as frequently as *bad*); using the same calculations as above, we would expect *good things happen* to occur ca. 1.23 times in the BoE. In fact, *good things happen* has a frequency of thirty-eight, which is almost thirty-one times more than expected. Therefore, even though the word form *good* is substantially more frequent than *bad* in the corpus — and we might, accordingly, expect *good* to collocate more frequently than *bad* with *things happen* — we observe *bad things happen* considerably more frequently in the corpus. It is difficult to generalize too much from this comparison because, as argued in Section 2.4.4, language is never random and so the import of expected frequencies is at very least subject to debate.

However, the fact that the word form *good* (369,849) is so much more frequent than *bad* (82,575) in the BoE in combination with the observation that the total raw frequency of the fifteen good collocates (1,025,140) vastly outnumbers the total of the fourteen bad collocates (247,407) in Table 7.2, is potentially revealing in the context of the current study. It could be suggested, for example, that the preference for negative collocates in this phrase, despite the underlying corpus frequencies is strong evidence for the phraseological nature of language itself. As Hunston argues, "[t]he semantic prosody of a lexical item is a consequence of the more general observation that meaning can be said to belong to whole phrases rather than to single words." Put another way, if meaning were created via the slot-and-filler open-choice model, we would almost certainly expect to find more instances of *good things happen* than *bad things happen* in the corpus. That this is not the case suggests that the completion of a phrase requires more than mere grammatical appropriateness.

Finally, it should also be recalled that the positive and negative lists used for automatic identification of evaluative collocates in Excel (see Section 5.5.1) are of very different sizes.

Recall that the negative list contains ca. 8,100 word forms, while the positive list contains only ca. 2000 word forms at the time of writing. This, combined with the fact that the raw frequencies of the good collocates in Table 7.2 so vastly outnumber those of the bad collocates, suggests that there are more negative words in English, but that users select the positive words more frequently. However, at this early stage of the research — the lists, as noted in Section 5.5.1 are not considered complete, and raw frequencies for these ca 10,000 items have not been retrieved — it would not be prudent to make such a strong argument. Future research would benefit from a more detailed exploration of this issue.

If we return now to the BoE T-Picture for *bad things HAPPEN* (69), an abridged version of which is reproduced in Table 7.3, we can see that both *bad things happen to good people* and *bad things happen to bad people* appear to be salient phrases (relevant collocates are shaded grey).

Table 7.3 Partial BoE T-Picture of *bad things HAPPEN*

N-1	NODE	N+1	N+2	N+3
when (7)	bad things happen (53)	to (24)	good(8)	people (13)
that (7)	happened (12)	in (9)	bad (3)	good (2)
why (5)	happening (4)	and (5)	people (3)	so (2)
where (4)		when (2)	war (2)	have (2)
or (3)		but (2)	other (2)	as (2)

Similarly, the BoE T-Picture for *good things HAPPEN* (38), an abridged version of which is reproduced in Table 7.4, indicates that *good things happen to good people* is salient (*good things happen to bad people* is also indicated by the picture, but *bad* (1) at N+2 is not shown in the table. These phrases are very long, and as such they occur very infrequently compared to the single wordform *happen* in the BoE. Recall that Sinclair (2003, p. 128) points out, for example, that "while *happen to be* is a phrase that is felt by native speakers to be quite normal and available, it is a lot less common than *happen to* and a great deal less common than just

happen," (italics added).

Table 7.4 Partial BoE T-Picture of *good things* HAPPEN indicating relevant phraseologies

N-2	NODE		N+1	N+2	N+3
many (3)	good things	happen (21)	in (3)	is (6)	first (3)
when (3)		happened (6)	to (3)	good (5)	people (3)
of (2)		happening (11)	you (4)	rowing (3)	good (2)
together (2)			for (3)	owe (3)	if (2)
see (2)			but (2)	fraser (2)	directing (1)

For this reason, BoE frequency data were checked against frequencies in the 19.6-billion-word English Web 2013[2] corpus (hereafter referred to as enTenTen13) to confirm that the low frequencies in the BoE data are not skewed in some way. Table 7.5 shows the evaluative polarity, raw frequencies, and normalized frequencies (expressed per million words of running text) of these phrases in both corpora.

Table 7.5 The phraseologies *good/bad things happen to good/bad people*: polarities, raw frequencies, and normalized frequencies in the BoE and enTenTen13 corpora

Phrase	Evaluative Polarity	BoE Freq	BoE Per Mil	enTenTen13 Freq	enTenTen13 Per Mil
things happen	Neutral	1,497	3.3	69,605	3.10
bad things happen	Negative	53	0.12	5,803	0.30
bad things happen to good people	Negative	7	0.02	1,002	0.04
bad things happen to bad people	Positive	2	0.00	48	0.00
good things happen	Positive	21	0.05	2,939	0.13
good things happen to good people	Positive	2	0.00	101	0.00
good things happen to bad people	Negative	1	0.00	61	0.00

As Table 7.5 shows, *bad things happen to good/bad people* accounts for nine (17%) of the fifty-three and 1,050 (18.1%) of 5,803 occurrences of *bad things happen* in the BoE and enTenTen13 respectively; these frequencies are, therefore, confirmed to be comparable. The phrase *good things happen to good/bad people* accounts for three (14.3%) of twenty-one and 162 (5.5%) of

2,939 occurrences of *good things happen*. Though these percentages are not as close as those of *bad things happen to good/bad people* in the two corpora, they are arguably not so dissimilar as to affect the current analysis.

The most remarkable effect of this extended phraseology in the context of the current study is that the addition of *to bad people* appears to reverse the evaluative polarity of *good/bad things happen*. This means simply that while *good things happen* is transparently positive on its own, *good things happen to bad people* is negative. Likewise, *bad things happen* is transparently negative, and *bad things happen to bad people* is positive. It is also important to note that in each case the phrases in Table 7.5 express evaluation via core evaluative meaning, not semantic prosody, which means that they require no additional context to reveal the evaluative meaning.

The collocational profile does seem to suggest that *things happen* is primed for evaluation, but this brief look at the effects of *to bad people* on the phrase *bad*/*good things happen* shows that neither the polarity nor the mode of evaluation are necessarily clear from the collocates lists alone.

7.1.2 Qualitative analysis of *things happen*

This section discusses results of the qualitative analysis of a random 100-line concordance of *things happen* (selected using the BoE built-in random line selection algorithm). The analysis began by removing twenty-three instances of the *MAKE things happen* and twenty-six of *these things happen*, because concordances for these phrases are the subject of detailed examinations in Sections 7.2 and 7.3. Table 7.6 shows the results of the analysis of the remaining fifty-one lines. The table shows that lines were labelled 'negative', 'positive', and 'neutral/unknown', and the negative and positive lines were further divided by whether evidence of semantic

prosody or textual evaluation was observed.

Table 7.6 Qualitative analysis of 51 lines of *things happen* in the BoE

Negative	Freq	%
Evidence of semantic prosody	9	17.6%
Evidence of textual evaluation	17	33.3%
Total	26	51.0%
Positive		
Evidence of semantic prosody	4	7.8%
Evidence of textual evaluation	2	3.9%
Total	6	11.8%
Neutral / Unknown	19	37.3%
Total	51	100.0%

Twenty-six (51.0%) of the lines express negative propositions, but only nine (17.6%) of these were considered instances of semantic prosody, in that the negative evaluation markers exhibit clear syntactic relationships with the node, as illustrated in line 1.

1. She also has her mobile phone, so when really hairy scary dreadful **things happen** such as, well, a lump of her tooth falling out while eating a peach, she is able to ring for a dentist.

Lines indicating textual evaluation contain linguistic or contextual information that indicates the polarity of *these things* but where there is no close syntactic relationship between these evaluation markers and *these things* or where extra-linguistic knowledge is required to activate the evaluation. Line 2 is an example of textual mode of evaluation because, even though *sombre* and the phrase *we don't like to see* clearly signal negativity, exactly what has happened remains unknown, even in this ca 200-character-long line.

2. described the mood at the airport yesterday as `sombre".`We don't like to see these sort of **things happen** at local airports," he said. `We have had a good safety record over the years and would like to see that

Overall, the qualitative study appears to contradict the quantitative data in at least one important way. The evaluative collocates suggest an almost equally positive and negative polarity, with

total FaC values skewing towards negative. The results of the qualitative study suggest that *things happen* is found in a negative environment much more often. However, it does not appear to be correct to claim that things happen has a negative semantic prosody when negative textual evaluation is observed in 33.3% of the lines, which is almost twice as many as the 17.6% that evince negative semantic prosody.

Also notable is that 19 lines contained no evidence of evaluation[3]. Since only larger contexts would reveal whether these lines evaluate positively or negatively, their mode of evaluation is likely textual. If we were to assume that this is the case for illustrative purposes, then almost 75% of the instances examined would be considered textual evaluation.

7.2 These Things Happen

This section and the next (7.3) investigate two phrases, *these things happen* (365) and *MAKE things happen* (366), which together account for 48.8% of the occurrences of *things happen* (1497) in the BoE. The previous section focussed on the adjective collocates at N-1 of *things happen*, but the profile of all collocates at N-1 of the T-Picture for *things happen*, reproduced in Table 7.7 below, shows the apparent salience — based on their very high frequencies — of the word form *these* and the lemma MAKE.

It is noteworthy that *these*, *MAKE*, *things*, and *happen* as individual words all have neutral core evaluative meanings, as do the phrases *these things happen* and *MAKE things happen*. These phraseologies are also, of course, examples of "nested collocations". As such, we can expect collocational and evaluative primings for *these things happen* and *MAKE things happen* to be different from each other and from the 'core' they share, *things happen*. In fact, as will be demonstrated, these phrases are primed to collocate and evaluate in very different ways.

Table 7.7 Top fifty collocates of *things happen* (1,497) at N-1 in the BoE, negative collocates in bold, positive collocates underlined

	Collocate	Freq		Collocate	Freq
1	these	365	26	odd	5
2	make	230	27	wonderful	5
3	making	77	28	interesting	5
4	**bad**	53	29	many	7
5	makes	47	30	great	6
6	strange	37	31	different	5
7	such	27	32	**dreadful**	4
8	two	27	33	weird	4
9	why	20	34	**awful**	4
10	**worse**	19	35	extraordinary	4
11	good	21	36	guess	4
12	those	19	37	see	6
13	when	22	38	if	9
14	**terrible**	15	39	several	4
15	way	16	40	important	4
16	let	14	41	dramatic	3
17	where	15	42	positive	3
18	funny	10	43	but	13
19	made	12	44	little	4
20	**horrible**	8	45	sometimes	3
21	how	10	46	some	6
22	night	8	47	special	3
23	whom	6	48	sure	3
24	certain	6	49	**irrational**	2
25	letting	5	50	unbelievable	2

7.2.1 Three Meanings of These Things Happen

The Longman Online Dictionary[4] entry for *these things happen* states that it is "used to tell someone not to worry about a mistake they have made, an accident they have caused etc." The MacMillan Dictionary[5] online defines *these things happen* in a subtly different way: it is "used for telling someone not to be upset about something unpleasant that has happened or something bad that they have done."

In fact, attested BoE occurrences show that *these things happen* (1,792) is used in at least three distinct ways. The least frequent meaning attested in the BoE is the literal, open-choice usage in which *these things* has specific, unambiguous referents in the text, as in line 3:

3. severe neglect as well as emotional, physical and sexual abuse. When **these things happen** to children, they internalise the experience and later, as a result, this may

The next most frequent sense is figurative. In these instances, the 'referent' of *things*, which may or may not be observable in the concordance line, is considered just one example of a type of thing that happens. In line 4, for example, the speaker is generalizing beyond what has happened to bemoan the fact that similar things seem to happen to her frequently.

4. They are spies." Oh, what have I got into, Violetta? Why do **these things happen** to me?" She paused and said musingly: `I think people who do not live according

Finally, the meaning closest to those expressed in the dictionary definitions above is the most frequent in the BoE. In these lines, *these things* also refers to "this type of thing", but there is an added element of offering comfort, consolation, or solace. In most cases, that which has happened is characterized as common/normal and therefore unworthy of undue concern, as illustrated in line 5.

5. The only disappointment is the goals we lost, but **these things happen** and we will be working hard to sort that out before the first competitive game."

Analysing three senses of *these things happen* separately is important because, as Hoey (2005, p. 13) argues, "[w]hen a word is polysemous, the collocations, semantic associations and colligations of one sense of the word differ from those of its other senses." For this reason, the 365 instances of *these things happen* in the BoE were divided into three concordances[6] after removal of five duplicate lines. The concordance containing instances of the open-choice iteration of the phrase contains only twelve lines; results of the analysis of these lines is discussed in Section 7.2.1.1.; discussion of the analysis of the sixty-line concordance in which *these things happen* is used to mean simply "this type of thing" follows in Section 7.2.1.2; and

analysis of the remaining 288 lines which were deemed to express the ameliorative meaning "mitigation of bad feelings" is presented in Section 7.2.1.3.

7.2.1.1 The open-choice phrase *these things happen*

No collocational profile was created for the twelve open-choice instances of *these things happen*. Because of the very small size of the concordance, all of the lines are reproduced in Figure 7.1. In these instances, the phrase appears to refer to literal referents, and most often at least one referent is observable in the *ca.* 200-character line[7]. Notably, though, very few would be observable in a standard collocational profile created from a 4:4 span or even a qualitative examination of the default 80-character lines presented by the BoE.

Lines 6 through 9 are negative and all contain direct referents to *these things*. They have, therefore, been labelled as indicators of negative semantic prosody. Lines 10, 11, and 12 also contain negative collocates, but there are no indications of what, specifically, happened. These lines have therefore been labelled indicative of negative textual evaluation. Lines 13 and 14 both contain indicators of positive evaluation. However, the first has been labelled prosodic, and the latter textual. Finally, lines 15, 16, and 17 contain no indication of either positive or negative evaluation, and so have been labelled 'unknown'.

It must be remembered that the fact that a referent is not visible in a concordance line does not mean that there is no referent in the text. In fact, the opposite is likely to be the case. Since general nouns, like *things*, are used as cohesive devices (Halliday and Hasan, 1976; Mahlberg, 2005), we can safely assume that the vast majority — even those considered textual here — contain observable referents in longer stretches of co-text. However, this returns us to the

theoretical issue that there is no clear delineation between semantic prosody and textual evaluation on the continuum of evaluative modes.

Figure 7.1 Twelve open-choice iterations of *these things happen* in the BoE

6. too was killed, Gyaltsen was arrested. When the nine-year-old saw **these things happen**,`I threw stones at the Chinese trucks, and one of my friends set fire to a jeep.
7. Are your fears realistic? Will doing the explorations really make **these things happen**? How could you protect yourselves against them?
8. severe neglect as well as emotional, physical and sexual abuse. When **these things happen** to children, they internalise the experience and later, as a result,
9. into ditches and generally fall to bits, but very rarely do any of **these things happen** in front of the cameras. So the highlights of the Top Gear Rally Reports
10. amount to confirming these allegations. It isn't true that I let **these things happen**, let alone knew about them." An Indonesian investigation team concluded
11. my first child was christened in the church, and to see all **these things happen** to it, it's just--it's just very heart rending.
12. said. Why not.7" Because it's the pool," Nick said. `That's where **these things happen**." But I first felt it grab me outside," Kelly said, in the flooded ground
13. as well as offering to relocate some of them from the region. Unless **these things happen** urgently, the stability of the area will be further threatened.
14. then to try to get everyone to coordinate and to cooperate to have **these things happen** sooner rather than later.
15. rumours Ershad will appoint a chief marshall administrator. And if **these things happen** that would mean his exceeding to the opposition's demand to hand over
16. some limited response from the Nato countries. Unless and until **these things happen**, it is right to ignore Belgrade's manoeuvres or to study them only as a
17. are thrown out of the guaranteed student loan program. Unless **these things happen**, saysfritz Elmendorf of the Consumers Bankers Association, banks will be

Line 11 above, considered an example of textual evaluation, exemplifies the problematic nature of seeking out direct referents in ever-expanding co-texts. The largest co-text viewable in the BoE[8] has been reproduced below. The ca. 200 characters initially examined are highlighted in grey for comparison:

> the blackened shell of the church. <p> Jim Painter: It was a church that was built by the people, not only financially but each piece of stick and stone that was put into it was done by the congregation. And it's such a tragic, tragic loss. It's just tough to come up to words when you're watching the church that

you--I was married in this church, my baby was--my first child was christened in the church, and to see all **these things happen** to it, it's just--it's just very heart rending. <p> Hosbein: Linda Goodwin is not only upset that the church building burned to the ground, she's concerned about the church's 22 preschool and day care employees who lost their jobs as a result of the tragedy. The recent fires have also spurred more than 100 worshipers to guard their churches around the clock. Some church leaders in north-central Florida, where most of the fires have

The text in the extended sample above arguably contains evidence of negative semantic prosody by definition; the negative referent is expressed twice in the passage. The first expression is found implicitly in the phrase *the blackened shell of the church*, at N-80 to N-75. The second is the explicit *burned to the ground* at N+22 to N+25. The difficulty lies in the fact that these structurally complex referents are so far from the node that we might reasonably question the semantic prosody label; they would never be observed in a standard collocational profile or a qualitative analysis of concordances shorter than 200 characters. The negative evaluation markers — *heart rending* and *upset* — found within the 200-character span evince textual evaluation because they are not the referents of *these things*.

Table 7.8 below summarizes the results of the qualitative analysis of the twelve open-choice iterations of *these things happen*. The table shows the number of lines which include evidence of negative and positive semantic prosody and textual evaluation as well as percentages of the total number of lines labelled negative and positive. The table shows the same calculations for lines labelled ‘unknown’.

Since there are so few lines, it is difficult to generalize from the data in Table 7.9. It does appear that the open-choice construction of the phrase *these things happen* is most often negative. However, it is not at all apparent that the negativity is expressed via positive-negative semantic prosody.

Table 7.8 Qualitative analysis of 12 open-choice iterations of the phrase *these things happen* in the BoE

Negative	Freq	%
Evidence of semantic prosody	4	33.3%
Evidence of textual evaluation	3	25.0%
Total	7	58.3%
Positive		
Evidence of semantic prosody	1	8.3%
Evidence of textual evaluation	1	8.3%
Total	2	16.7%
Neutral / Unknown	3	25.0%
Total	12	100.0%

7.2.1.2 *These things happen* meaning "this type of thing happens"

This section reports on analysis of the sixty-line concordance containing *these things happen* in which *these things* means "this type of thing." A collocational profile was created by importing the sixty lines from the BoE into Microsoft Excel where a 4:4 Positional Frequency Table (PFT) was created. Because of the small number of lines in the concordance, some of the columns in the table do not reach the (arbitrary) cut-off of fifty collocates. This potentially serious problem is examined in more detail in Section 7.2.1.3.

Table 7.9 shows that very few collocates in the profile are evaluative. Note that FaC values are not labelled in Table 7.9 because each occurs only once. Although negative collocates do outnumber positive in the standard 4:4 span by a ratio of 2:1 (12 negative and 6 positive) there are not nearly enough to support a persuasive argument for negative semantic prosody. Additionally, while the exact relationship between collocates and node remains unknown, only four collocates in Table 7.9 could be specific 'referents' of *these things — overreaction, shock, trouble, problems.*

T-scores are not included in Table 7.9 because the PFT from which the evaluative collocates

were extracted was created in Excel. As noted in Section 5.4.2, it is technically possible to use Excel to calculate t-scores using, for example, Barnbrook's (1996, p. 97) formula, but bespoke calculations of this kind are difficult to execute in Excel and are beyond the scope of this thesis. It is worth noting, however, that with FaC values of one in each case, the collocates in Table 7.9 are likely to have t-scores below the standard level of 2.0.

Table 7.9 Negative and positive collocates in the 4:4 Positional Frequency Table for the sixty lines of *these things happen* meaning "this type of thing"

Negative								
N-4	N-3	N-2	N-1	NODE	N+1	N+2	N+3	N+4
overreaction	unfortunately	died	—		—	—	—	problems
appalling		shock						disappointed
wrong		trouble						killers
		uncomfortable						
		unfair						
Positive								
—	benefits	extraordinary	—		—	—	important	—
	help	funny					peace	

To illustrate the potential problem, Table 7.10 shows collocates with t-scores higher than 2.0 from N-1 to N-4 for the full concordance of *these things happen* (360) in the BoE. The t-score cut-off applied to collocates of *things happen* (see Table 7.1 above) removes a roughly equal number of positive and negative collocates and therefore does not seem to substantially affect the overall ratio.

However, the same method applied to collocates in Table 7.10 leaves us with no evaluative collocates at all, when, in fact, there are as many as twenty-six negative and eight positive collocates in the top-fifty at N-1 to N-4 of the full concordance of *these things happen* (360).

Table 7.10 Collocates of *these things happen* (360) at N-1 to N-4 in the BoE, showing only collocates with t-scores higher than 2.0

	N-4			N-3			N-2			N-1		
	Collocate	FaC	T-score	Collocate	FaC	T-score	Collocate	FaC	T-score	Collocate	FaC	T-score
1	it	16	3.18	very	7	2.51	he	11	2.61	but	66	7.91
2	a	17	2.16	or	8	2.44	why	5	2.16	said	18	4.02
3				i	10	2.29	say	5	2.12	that	20	3.64
4										when	13	3.38
5										well	10	3.05
6										how	7	2.51
7										<p>	13	2.48
8										sometimes	6	2.42
9										why	6	2.38
10										know	6	2.30
11										moment	5	2.20
12										make	5	2.10

A qualitative approach, however, results in a very different account of how these instances evaluate. Table 7.11 shows the total number of negative and positive lines and which of these includes evidence of semantic prosody or textual evaluation.

Table 7.11 Qualitative analysis of 60 instance of *these things happen* meaning "this type of thing" in the BoE

Negative	Freq	%
Evidence of semantic prosody	6	10.0%
Evidence of textual evaluation	24	40.0%
Total	30	50.0%
Positive		
Evidence of semantic prosody	3	5.0%
Evidence of textual evaluation	11	18.3%
Total	14	23.3%
Neutral / Unknown	16	26.7%
Total	60	100.0%

As Table 7.11 shows, half of the lines appear to evaluate negatively, but only 10% show direct evidence of negative semantic prosody in the form of a word or phrase that can be characterized as "the type of thing" that is being referred to. By comparison, 40% of the lines contain items that evaluate negatively and seem to indicate a more general, textual evaluation but are not in

themselves syntactically related (as referent to the general noun) to *these things*. Less than one quarter of the lines appear to evaluate positively, and only 5% contain observable evidence of positive semantic prosody.

Table 7.11 also shows that sixteen of the sixty lines (26.7%) have been labelled "unknown". As suggested previously, it is likely that larger co-texts would reveal the polarities of these lines and the type of evaluation they exhibit. However, the fact that, even in these relatively long lines, just over one quarter remain unlabelled is itself evidence of the potentially problematic nature of observing modes of evaluation in studies of positive-negative semantic prosody.

7.2.1.3 *These things happen* used to mitigate negativity

The most frequent usage of the phrase *these things happen* appears, unsurprisingly, to be the one cited in the Longman and MacMillan dictionary entries above. This iteration, like the one discussed above, is a semi-preconstructed phrase and is therefore selected whole, and not generated word by word. It is tempting to claim that *these things happen*, as it is used in relation to feelings of guilt, sadness, disappointment, etc., has a negative core meaning, i.e. the things that happen are by their very nature unfavourable. Even when used with humorous intent, the referent is still transparently negative, as in line 18:

18. In the office there's a phone call from a woman whose cat has clawed her prosthesis into shreds. `**These things happen,"** she says Lesley, trying not to laugh.

However, the fact that the phrase is used to mitigate or assuage these feelings is problematic. In this sense it is similar to the examples provided by Whitsitt (2005) (see Section 1.3.1 above for a detailed discussion of *alleviate, heal, relieve*, etc.). In answer to Whitsitt, Morley and Partington (2009, p. 142) employ a specialized notation to express what they call the lexical

item's "embedded evaluation". In instances like line 18, however, no simple version of their notation seems possible. Whereas the notations "[*exacerbate* [a problem]] and (*alleviate* [a problem])" (Morley and Partington, 2009, p. 142) exemplify an elegant response to Whitsitt's objection, similar notation of the example in line 18 — i.e. (*these things happen* [cat has clawed her prosthesis into shreds]) — is a clumsy approximation, primarily because there is an added layer of meaning. The speaker is "writing the device" (Louw, 1993) by using the phrase ironically, and is, notably, *trying not to laugh*. Morley and Partington's notation does not appear to account for ironic uses of evaluative items.

Table 7.12 shows both positive and negative collocates and their FaCs in the 4:0 span of the 288-line concordance of "mitigating negativity" meaning of *these things happen*. The evaluative collocates in the 0:4 span are not presented here because there are too few to warrant specific comment, although it is worth noting that negative outnumber positive 6:2. Although negative collocates outnumber positive 2:1 (14:7) in Table 7.12, the total numbers are not convincing evidence of negative semantic prosody. Specifically, these fourteen negative collocates represent only 7% of the 200 collocates in the top-fifty at N-1 to N-4. Additionally, there are indications of two potentially serious problems with the data in Table 7.12. First, as Stubbs (2007, p. 171) notes, "[f]requency cut-off points are clearly arbitrary, but can always be lowered to give a more delicate description." In this case that a threshold of even two occurrences would reduce the total number of negative collocates to six (3% of the 200 total collocates in the PFT) and leave only four positive (2% of the 200). There would remain three evaluative collocates (all negative: *unfortunately*, *mistake*, *killer*) at N-1; five at N-2 (three negative: *unfortunate*, *disappointing*, *sad*; two positive: *free*, *fans*); and one positive at each at N-3 and N-4 (*good*, *passion*).

Table 7.12 Evaluative collocates in the 4:0 span of the 288-line concordance of the phrase *these things happen*; extracted from the top-fifty PFT created in Excel

Negative							
N-4	FaC	N-3	FaC	N-2	FaC	N-1	FaC
overwhelming	1	blame	1	unfortunate	4	unfortunately	2
lose	1			disappointing	3	mistake	2
criticise	1			sad	2	killer	2
shame	1			terrifying	1	blame	1
						worry	1
Total	4		1		10		8
Positive							
passion	2	good	2	free	2	pleasant	1
appealing	1	realistic	1	fans	2		
Total	3		3		4		1

In itself, this is not overly problematic. The human analyst can employ the additional required 'delicacy' and apply the lower frequency threshold if appropriate. However, this leads to another potential problem in creating and analysing PFTs. Specifically, in any list ordered by frequency, collocates with the same FaC value will be ordered pseudo-randomly (based on the order in which they appear in the concordance). Again, in itself this is not normally a problem, but when the number of collocates occurring only once extends beyond the top-fifty allowed by the BoE software or Excel macros used to create PFTs it is impossible to know if the observable evaluative collocates are representative of the overall ratio.

Generally, the lines copied from the BoE and pasted into Excel were unsorted and therefore preserved the order in which they were originally presented by the BoE telnet interface. This means that at N-1 of Table 7.12 the negative collocates *blame* and *worry* appear only because they occur earlier in the concordance. If the lines were resorted and a new table created, different collocates would be shown in these slots. In fact, at N-1 of the 288-line concordance 118 collocates occur only once. Table 7.13 shows that twenty of these are negative and three are positive.

Table 7.13 Evaluative collocates in the 4:0 Span of the 288-line concordance of *these things happen* in the BoE

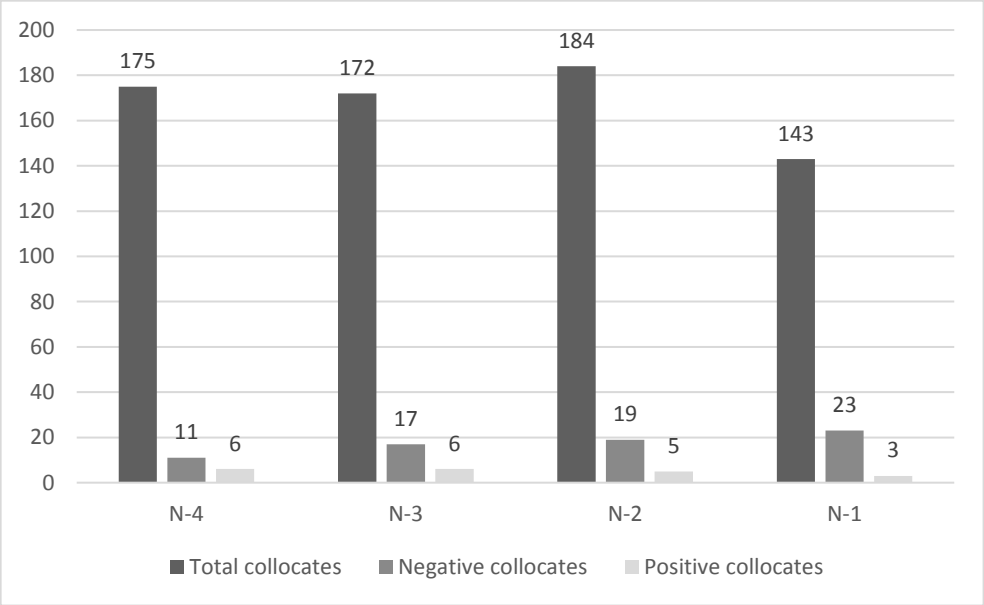
Negative							
N-4	FaC	N-3	FaC	N-2	FaC	N-1	FaC
overwhelming	1	blame	1	unfortunate	4	unfortunately	2
lose	1	missing	1	disappointing	3	mistake	2
criticise	1	bad	1	sad	2	killer	2
shame	1	shatter	1	terrifying	1	blame	1
hell	1	terribly	1	lost	1	worry	1
bugger	1	incident	1	disappointed	1	afraid	1
diseases	1	mistakes	1	shocked	1	incident	1
concerned	1	despair	1	embarrassing	1	fumed	1
bruising	1	shame	1	error	1	mistakes	1
blaming	1	irritable	1	bust	1	injuries	1
outraged	1	broken	1	unreasonable	1	oops	1
		defeat	1	spluttered	1	panic	1
		worry	1	suffer	1	clash	1
		sorry	1	ignore	1	disagreement	1
		damage	1	terrible	1	abuse	1
		collision	1	discordant	1	excuses	1
		prison	1	trouble	1	piss	1
				regrettable	1	sorry	1
				missed	1	despondent	1
						unfair	1
						aberration	1
						downhearted	1
						consternation	1
Total	11		17		19		26

Positive							
passion	2	good	2	free	2	pleasant	1
appealing	1	realistic	1	fans	2	confidence	1
sensible	1	great	1	better	1	important	1
incredible	1	fine	1	good	1		
spiritual	1	wonderful	1	unabashed	1		
masterpiece	1	revitalise	1				
Total	7		7		7		3

When the table shows all evaluative collocates in the 4:0 span, there appears to be a much stronger indication of negative semantic prosody. The ratios of negative to positive collocates at N-1 is 26:3; at N-2 it is 19:7; at N-3 it is 17:7; and at N-4 it is 11:7 (for a total of 73 negative to 24 positive). Despite these potentially convincing ratios, however, the graph in Figure 7.2 illustrates that it is difficult to sustain a convincing argument that *these things happen* has a negative semantic prosody based on the numbers of negative to positive collocates. As the graph

shows, there are 143 unique collocates at N-1. Of these, only twenty-three (16.1%) are negative. At N-2, only nineteen (10.3%) of the 184 unique collocates are negative; at N-3 and N-4 only 9.9% and 6.3% of the unique collocates are negative. These are hardly convincing percentages.

Figure 7.2 Graph comparing total number of collocates to negative and positive collocates at N-1 to N-4 of the 288 BoE lines of *these things happen*



The total FaC values are very similarly sparse. As we have seen, it is possible for a small number of evaluative collocates to account for a very large number of occurrences in the corpus. But that is not the case for *these things happen*. For example, the twenty-three negative collocates at N-1, for example, account for only twenty-six (9.0%) of the 288 lines in the concordance. Similar very small evaluative FaC totals are found at N-2 to N-4, as shown in Figure 7.2 above.

It is also worth noting that T-scores are not shown in the table. This is because none of the evaluative collocates of *these things happen* in the 4:4 span have t-scores higher than 2.0. The highest are *unfortunate* at N-2 and *passion* at N-4, the t-scores of which are both 1.99. Of course, this means that if statistical significance were a requirement of evaluative collocates, none at all would be observable for this phrase in the BoE.

For the qualitative analysis, 100 of the 288 lines were selected using Microsoft Excel's built-in random number generator. Results of the qualitative study are shown in Table 7.14.

Table 7.14 Qualitative analysis of 100 lines of *these things happen* as it is used to mitigate/assuage negative feelings

Negative	Freq	%
Evidence of semantic prosody	27	27.0%
Evidence of textual evaluation	56	56.0%
Total	83	83.0%
Positive		
Evidence of semantic prosody	3	3.0%
Evidence of textual evaluation	6	6.0%
Total	9	9.0%
Neutral / Unknown	8	8.0%
Total	100	100.0%

As Table 7.14 shows, 83% of the lines contain evidence of negative evaluation, and only 9% are positive. However, only 27% were considered to evaluate via semantic prosody. Lines 19 and 20 below illustrate:

19. groaned: `It was a disappointing way to lose the game, but sometimes **these things happen**, you just have to accept it and cannot change results." Dr Jo is in desperate know ...
20. 37, puts the row down to a clash of personalities. <p> He said: `**These things happen** all the time. <p> When I played, people argued with the manager or coach all the

The remaining negative lines, 56% of the total, were labelled negative textual evaluation as in lines 21 and 22:

21. in the most ardent `no fault" advocate's car while I mumble: `Sorry, **these things happen**, let's split the cost 50-50." The response would be explosive outrage. Marriage
22. victim. <p> Initially, one felt sorry for both sides. After all, **these things happen**, nobody's perfect, blah, blah, blah. But when Cruise announced the new before

Eight lines were labelled 'unknown' because there is no observable evidence of evaluation of any kind, as in 23 and 24 below:

```
23.          to me, `New Hampshire is using to build roads with," adding,
   `These things happen." But the chairman was also getting the word that
   Hillary Rodham Clinton wanted
24.        weekend telling all the players it was simply a one-off and that
   these things happen in football." <p> Meanwhile, Andy Webster, the
   Hearts youngster, insisted
```

Again, the fact that in this amount of context there is no apparent evaluation observable in lines 23 and 24 (and seven more lines in the concordance) does not mean that the phrase does not evaluate in these instances. Rather, it is simply the case that neither the direct referent nor any textual clues to the evaluative polarity are observable in the span of text available.

7.3 MAKE Things Happen

This section reports on the investigation of the 366 occurrences of *MAKE things happen* in the BoE. As noted earlier in the chapter, the lemma MAKE initially appeared to be a salient collocate when it was observed to account at N-1 for 366 (24.4%) of the 1,497 occurrences of *things happen* in the BoE. The analysis began by removing twenty-five lines, leaving 341 for analysis: duplicates (3); lines where the phrase is used as a heading (14) or sub-heading (4)[9]; the coincidental co-selection "[…] made. Things happen […]" (1); and open-choice (non-figurative) variants (3).

7.3.1 Meaning of MAKE things happen

There are no dictionary definitions for "make things happen" in the Collins, MacMillan, or Longman online dictionaries but close readings of attested examples indicate that the phrase is used to characterize a person or group who creates or finds opportunities to do something beneficial, usually in a very specific arena such as sport or business. These beneficial things are very rarely named, however. Instead the phrase refers to a general, indeterminate set of contextually relevant actions/behaviours. For example, in line 25 what Voss makes happen is

not specified (nor is there any specific referent in the extended co-text), but he is presumably already known to the reader either via previous contextual knowledge or has been introduced previously in the article.

25. Voss became one of the AFL's elite by **making things happen** on the field immediately, not in some vague netherworld where progress is slow

This shared knowledge in combination with the mention of the "AFL" and the fact his actions *happen on the field* communicate to the reader a clear understanding of what types of things are made to happen.

7.3.2 Collocational profile of MAKE things happen

This section explores the collocational profile of the lexical phrase *MAKE things happen*. Since it was noted above that Pictures and PFTs (especially those created from relatively small concordances) may omit potentially relevant collocates based on arbitrary ordering of concordance lines, a 4:4 PFT was created that ignore the arbitrary 50-row limit. Table 7.15 shows the evaluative collocates in the 4:0 span; and Table 7.16 shows the evaluative collocates in the 0:4 span.

Like the collocates of *these things happen* above, the vast majority of the evaluative collocates of MAKE *things happen* occur only once each at each position in the span. Once again, t-scores are not included in the tables in this section, but it should be noted that none of the scores are above the usual significance threshold of 2.0. If we combine numbers of evaluative collocates in Table 7.15 and Table 7.16, the total ratio of positive to negative is just less than two to one (57:33), but the numbers of evaluative collocates as a percentage of the total number of collocates in each slot again seem to belie the raw ratio. The highest number of positive

collocates in Table 7.15 — nineteen — is found at N-2, but these represent only 8.8% of the total of 217 unique collocates in this position (see Figure 7.3 below).

Table 7.15 Evaluative collocates and their FaCs in the 4:0 span for the phrase MAKE *things happen* in the BoE

Negative							
N-4	FaC	N-3	FaC	N-2	FaC	N-1	FaC
moronic	1	dangerous	1	aggressively	1	upsets	1
tantrum	1	ruthless	1	problem	1	aggression	1
scrap	1	cliches	1	attack	1		
desperately	1	problems	1	noise	1		
mere	1	weakness	1	failed	1		
scrap	1			rattle	1		
problems	1			risks	1		
				loose	1		
Total FaC	7		5		8		2

Positive							
enthuse	1	tremendous	2	enthusiasm	1	hale	1
flexibility	1	good	2	love	1	enjoy	1
professional	1	great	2	good	1		
unobtrusively	1	creative	1	laureate	1		
revitalise	1	decisive	1	promised	1		
entertain	1	celebrates	1	effective	1		
winners	1	awards	1	fun	1		
energetic	1	top	1	win	1		
enjoy	1	innovative	1	expertise	1		
fun	1	respect	1	lucky	1		
helped	1	expertise	1	nice	1		
		talented	1	commitment	1		
				elite	1		
				best	1		
				destiny	1		
				dynamic	1		
				talent	1		
				eager	1		
				skill	1		
Total FaC	11		15		19		2

Similarly, Table 7.16 shows that there are eighteen positive collocates at N+4, representing 8.2% of the 219 unique collocates of MAKE *things happen* in this position. Even without taking into account the low FaC values of these collocates, these percentages are too low to support a strong

argument for the positive semantic prosody of MAKE *things happen* in the BoE.

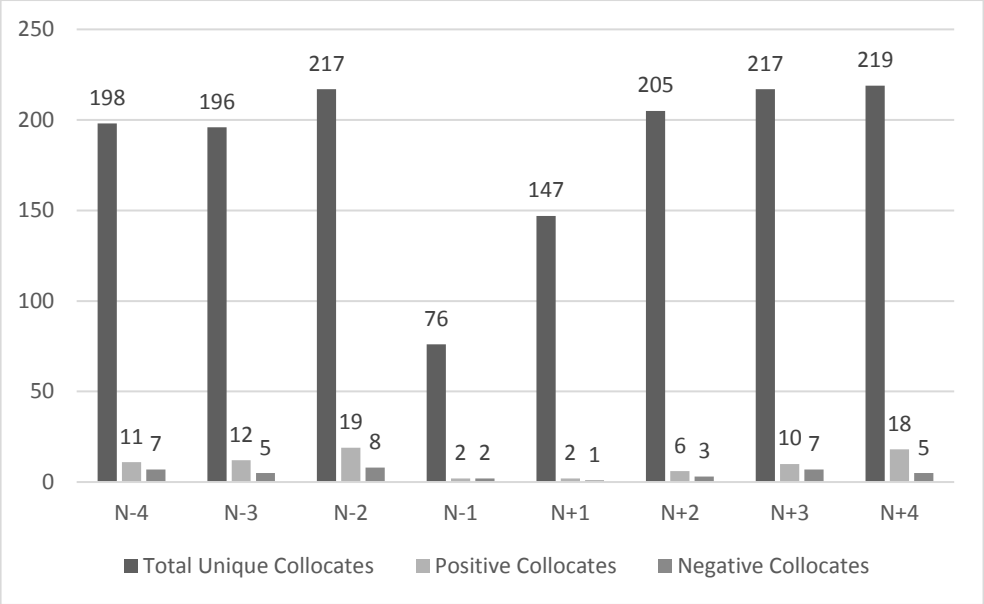
Table 7.16 Evaluative collocates and their FaCs in the 0:4 span for the phrase MAKE *things happen* in the BoE

Negative							
N+1	FaC	N+2	FaC	N+3	FaC	N+4	FaC
apocalypse	1	worry	1	risk	1	atrophy	1
		fear	1	inadequacy	1	failed	1
		difficult	1	spite	1	missing	1
				defeats	1	lied	1
				delay	1	death	1
				lurked	1		
				difficult	1		
Total FaC	1		3		7		5
Positive							
good	1	good	1	important	2	strong	2
faith	1	friends	1	magic	1	good	2
		amazing	1	excited	1	concise	1
		intelligent	1	motivated	1	victory	1
				wonderfully	1	defender	1
				prompt	1	gifted	1
				revolutionized	1	interesting	1
						awe	1
						help	1
						wise	1
						safe	1
						capable	1
						adventurous	1
						loved	1
						award	1
						important	1
						favour	1
						energetic	1
Total FaC	2		4		8		20

The combined totals from the two tables are presented in the graph in Figure 7.3 below. As with *these things happen* in the previous section, Figure 7.3 clearly illustrates that despite what at first appears to be a convincing positive to negative ratio, the total numbers of evaluative collocates are too low to convincingly support an argument for semantic prosody. If we take into account the FaC values, the evidence is even less convincing. At N-2, the Total positive FaC is also nineteen (each collocate occurs only once in this position), which accounts for only

5.2% of the 366 occurrences of the phrase. Likewise, the total positive FaC of twenty at N+4 accounts for only 5.5% of occurrences.

Figure 7.3 Total unique collocates, positive collocates, and negative collocates in the 4:4 span of the 341-line BoE concordance of MAKE *things happen*



7.3.3 Qualitative analysis of MAKE things happen

As in the previous section, 100 lines were selected randomly for qualitative analysis. First, evidence of positive and negative evaluation was noted in each line. However, clear and direct referents with close logical relationships to the node, which are thus considered evidence of semantic prosody, are very rarely observable. At most, what is observable in the lines tends to be further general or even quite vague descriptions of what is made to happen, not an explicit expression.

Figure 7.4 contains all eight lines of the 100 lines deemed to contain evidence of semantic prosody. One of the clearest indicators of prosodic evaluation is found in line 26, in which the

phrase *Showed good instinct to score the goal* is a concrete example of one of the specific positive things the player makes happen. Only two lines, 32 and 33, were labelled negative. The remaining six express positive propositions.

Figure 7.4 Eight lines containing evidence of semantic prosody in the 100-line BoE concordance of *MAKE things happen*

26. Aodhan macgearailt 7 INDUSTRIOUS and always looking to **make things happen**. Showed good instinct to score the goal.
27. and foremost a catalyst rather than a principal. He orchestrates, he **makes things happen**. His job is to make the various relationships work:
28. some might feel the need to set the world alight, others are content to **make things happen** in small ways. The sense of self'-fulfilment that you can get from giving pleasure to others and
29. Craig Brown to solve. Lambert is a big player for his country. He **makes things happen** for them in the same way as for Celtic. He makes teams tick, always covering, always getting in
30. sheet," said Tommy Taylor, Orient's manager. `The forwards tried to **make things happen** and we put a few crosses in, although they weren't quite good enough.
31. He questioned the value of poetry throughout his life '- could it really **make things happen**? Did it contribute to the good of mankind?
32. prime mission to pursue and arrest him. She is a ruthless achiever who **makes things happen** by force of personality.
33. the degradation of women. <p> Because it is an activity where the press of a key can **make things happen**, it's time to think of porn as a visit to a red'-light district,

Most lines in the concordance examined appear to evaluate positively via textual meaning, as in 34 below where *valued* and *which are important to us* indicate positive evaluation, but *things* has no observable referent or even any indication of what kind of thing it refers to. Similarly, line 35 contains multiple expressions of positive evaluation, but no indication at all of the referent might be.

34. are valued instrumentally, of course. But so, too, are human beings-they **make things happen** which are important to us.
35. It was a fantastic feeling to be out there **making things happen**. The reaction of the crowd was just like the good old days. They were right behind me

Table 7.17 below summarizes the results of the qualitative analysis of the 100-line concordance of *MAKE things happen*. The table illustrates that not only does this phrase evaluate

overwhelmingly positively, but that evaluation is also mainly textual, with 83% of the lines displaying evidence of evaluation that requires extra linguistic or cultural knowledge for the evaluation to be observed.

Table 7.17 Results of qualitative analysis of 100 BoE lines of *MAKE things happen*

Negative	Freq	%
Evidence of semantic prosody	2	2.0%
Evidence of textual evaluation	3	3.0%
Total	5	5.0%
Positive		
Evidence of semantic prosody	6	6.0%
Evidence of textual evaluation	83	83.0%
Total	89	89.0%
Neutral / Unknown	6	6.0%
Total	100	100.0%

7.4 Conclusion

Data presented in this chapter show that phraseological behaviour can have considerable effects on how an item evaluates. First, it was shown that although the nested collocation *things happen* has nearly the same number of positive and negative adjective collocates at N-1, the summed FaC values show that negative collocates are selected *ca.* 30% more often. Further, it has been demonstrated that the 'addition' of the phrase *to bad people* reverses the polarity of the core evaluative meanings of *bad*/*good things happen*.

Collocational analysis of the 4:0 span of the 288-line concordance for *these things happen* illustrates several potentially serious issues. First, although the ratio of evaluative collocates in the 'top fifty' list was 2:1 (14 negative to 7 positive), these raw numbers were once more deemed too low to make any strong argument for negative semantic prosody. When the 'arbitrary' ordering of collocates is avoided and all are accounted for (not just the 'top' fifty),

we find the ratio is closer to 3:1 (73 negative to 24 positive). However, the numbers of unique negative collocates as percentages of the total unique collocates in each position from N-1 to N-4 were, again, much too small to indicate negative semantic prosody. The highest percentage of negative collocates is 16.1% (or 23 out of 143) at N-1.

Closely related are the FaC totals which suggest that the pragmatic decision to evaluate negatively is not observable in strictly collocational evidence. The twenty-three negative collocates at N-1, for example, are found in only twenty-six (9.0%) of the 288 lines of *these things happen* in the BoE. It was also noted that if the standard rules of collocational significance were required of the collocates examined (in this case a t-score of 2.0 or above) we would be left no evidence for consideration.

Qualitative examination of the twelve lines of the open-choice variant of *these things happen* shows an overall tendency toward negativity (seven of twelve lines), but the modes of evaluation were almost equally split between prosody (four lines) and textual evaluation (three lines). Examination of a larger concordance of open-choice instances would be needed to establish whether this split is indeed indicative of how *these things happen* evaluates or if one of the modes is in fact more frequent. Qualitative investigation of the 60 lines meaning "this type of thing" revealed that 50% of the lines evaluated negatively, but this percentage was 10% prosodic and 40% textual. More than 25% of the lines were left 'unknown' and it is hypothesized that extended co-text would reveal that many, if not most, of these instances evaluate via negative textual meaning. Finally, the qualitative examination of 100 lines of the ameliorative meaning of *these things happen* reveals that 83% of the instances do appear to evaluate negatively, but as the low percentages of negative collocates suggests, relatively few of these evaluate via semantic prosody. Only 27% are considered to contain evidence of

semantic prosody, and 56% are deemed negative only by applying linguistic or contextual knowledge.

The only major difference between the analyses of *these things happen* and *MAKE things happen* is that the evaluative polarity of the latter is overwhelmingly positive. The collocational profile for *MAKE things happen* once again contains too few evaluative collocates even when the search extends beyond the 'top' fifty; the highest percentage of positive collocates in the 4:4 span is only 8.8% at N-2. The qualitative analysis demonstrates that, although overwhelmingly positive, the vast majority of lines evaluate via textual meaning. Only 6% contained evidence of semantic prosody.

Notes

1 The BoE query JJ+things+happen@ (any word tagged as an adjective + *things* + the lemma HAPPEN returns 637 lines. To ensure the adjectives returned were in fact modifying *things* and were not occurring in that position by coincidence, the concordance was 'cleaned' by removing instances of *Things happen* and instances where *things happen* is preceded by punctuation.

2 https://the.sketchengine.co.uk, accessed on 3 January 2017.

3 It is also worth noting that things happen exhibits a clear preference for 'strange' or 'unexpected' adjective collocates at N-1. These collocates are neither explicitly positive nor negative in themselves, and even in larger contexts it is often difficult to determine if or how these adjectives contribute to evaluation.

4 http://www.ldoceonline.com/dictionary/these-things-happen, accessed 26 December 2016

5 http://www.macmillandictionary.com/us/dictionary/american/these-things-happen, accessed 12 January 2017

6 It is not always possible to label the lines definitively. The concordance was divided based on my own impressions/intuition and often represent little more than a best guess. A larger context or different interpretation by a different researcher could result in slightly different frequencies. However, these differences are not likely to be large enough to significantly affect the results of this study.

7 As noted in Section 5.4.3, the 200-character lines are often edited into shorter versions for the purposes of formatting and increased readability.

8 Subtle manipulations of search queries, etc., can reveal more text to either side, but never all at once. This section of text represents the standard method and amount of expansion. In my experience the command that allows a user to view the full text rarely works, presumably due to copyright restrictions.

9 Headings and sub-headings are marked in the BoE with the <h> and <subh> tags respectively. They were removed because they do not have a direct syntactic relationship with the 'line' they appear in.

CHAPTER 8:
FURTHER EFFECTS OF PHRASEOLOGICAL BEHAVIOUR ON THE SEMANTIC PROSODY OF HAPPEN

8.1 Introduction

The chapter is divided into two main sections. Section 8.2 focusses on *a\an NOUN waiting to happen* and 8.3 focusses on *the ADJ thing that can happen*. These two main sections are in turn divided into subsections comprising discussions of how phraseology specifically affects observations of semantic prosody in corpus data. In each section, discussion first focusses on collocational profiles of various iterations of the phrases (e.g. *happen, to happen, waiting to happen, accident|disaster waiting to happen*); these "nested combinations" (Hoey, 2005) are expected to produce different collocational profiles, of course, but the sizable differences in numbers and FaC values for the evaluative collocates in the profiles was unexpected.

Discussion then returns to the relationship between collocation and semantic prosody and, specifically, the role phraseology plays in the observation of evaluative collocates. It is argued that many apparently significant evaluative 'collocates' observed in collocational profiles of HAPPEN and its word forms are observed primarily because they are elements of frequent phrases. It is argued, therefore, that such collocates contribute only superficial evidence of semantic prosody.

This leads to the observation that these two semi-preconstructed phrases evaluate via their core meanings. The evaluative polarity of these core meanings appears to be an effect of "internal lexical variation" (Sinclair, 1991, p. 111), namely each phrase has at least one variable word, the core polarity of which appears to determine the meaning of the phrase as a whole. It is also argued in Section 8.3.4.3 that the general noun *thing* plays an important role in the evaluative

nature of *the ADJ thing that can happen*, and this observation supports the hypothesis that the entire phrase is primed to evaluate irrespective of which collocate is selected.

Finally, qualitative analyses of concordances are discussed. Specifically, it is noted that there does not appear to be anything in the immediate context of the node (in the *ca.* 200-character lines examined) affecting the evaluative polarity or mode of either of these phrases.

8.2 A|An NOUN Waiting to Happen

The phrase *a|an NOUN waiting to happen* was selected for this study because it was observed in an early pilot study of MWUs. It is not included in the MWUs above (Table 5.4) but it is clearly apparent in BoE Pictures of *to happen* in which *waiting* is the sixth most frequent collocate at N-1.

8.2.1 Collocational profiles of phraseological iterations of *NOUN waiting to happen*

This section discusses collocational profiles at each stage in the 'growth' of the phrase *accident|disaster waiting to happen*. The tables presented below show only evaluative collocates in the 4:0 span because this is where the syntactically relevant collocates (i.e. grammatical subjects of the verb happen, and so on) are more likely to be found. The varying numbers and frequencies of evaluative collocates in the 4:0 span of the T-Pictures for *happen*, *to happen*, *waiting to happen*, and *accident|disaster waiting to happen*, demonstrate that although all iterations do show at least some evidence of negative semantic prosody, only *waiting to happen* is strongly primed for negative semantic prosody. The implications of this observation are discussed in detail in the sections that follow.

First, the T-Picture (4:0 span) for the word form *happen* contains relatively few indications of

negative semantic prosody, as Table 8.1 shows. While there are in fact more negative collocates than positive, neither their numbers nor their raw FaCs (individually or combined) are entirely convincing. Even *worst* and *best* at N-4 (shaded in the table), which have the highest FaC values of the evaluative collocates, do not necessarily evince the negative semantic prosody of the lone item *happen*. This is because they appear as 'collocates' in this position almost entirely due to their occurrences in the phrases *the worst|best thing that can happen*.

Table 8.1 Evaluative collocates in the 4:0 span of the T-Picture for *happen* (43,759) in the BoE

Negative												
N-4			N-3			N-2			N-1			
Collocate	FaC	T-score	Collocate	FaC	T-score	Collocate	FaC	T-score	Collocate	FaC	T-score	
worst	155	12.18	worst	124	10.84	accidents	116	10.71	accidents	89	9.37	
fear	58	7.00	accident	87	9.09	bad	104	9.40	mistakes	18	4.01	
worried	45	6.36	disaster	83	8.91							
bad	54	6.25										
Total	312		Total	294		Total	220		Total	107		

Positive												
Collocate	FaC	T-score	Collocate	FaC	T-score	Collocate	FaC	T-score	Collocate	FaC	T-score	
best	83	6.70	—	—	—	—	—	—	miracles	19	4.31	

Recall that in each 'slot' (N-1, N-2, etc.) the BoE software lists fifty collocates. Therefore, the two negative collocates at N-1 of Table 8.1, for example, represent only 4% of those shown by the software. Perhaps more noteworthy is the fact that the frequencies are quite low. In the T-Picture at N-1, *accident* and *mistakes* are ranked twenty-ninth and forty-ninth respectively, accounting for only 107 instances combined. This represents only 0.24% of the 43,759 instances of *happen* in the BoE.

The relative scarcity of evaluative collocates in this profile is not especially surprising. As discussed in Chapter 6, evidence for the semantic prosody of the 'bare' lemma HAPPEN is likewise quite sparse. Evidence of the negative semantic prosody of HAPPEN is strongest in a

collocational profile that contains only nouns in the 4:0 span.

Table 8.2 below also contains very few evaluative collocates, and, again, combined frequencies are relatively low when viewed as percentages of the total occurrences of the phrase. Even at N-2 in Table 8.2, where the majority of negative collocates are found, there are only five totalling 208 occurrences; this means that of the 10,938 instances of *to happen* in the BoE, only 1.9% have negative collocates in this position. It is difficult to support a strong argument for negative semantic prosody on such a low percentage.

Table 8.2 Evaluative collocates in the 4:0 span of the T-Picture for *to happen* (10,938) in the BoE

Negative												
N-4			N-3			N-2			N-1		
Collocate	FaC	T-score	Collocate	FaC	T-score	Collocate	FaC	T-score	Collocate	FaC	T-score
—	—	—	bad	37	5.75	accident	77	8.71	failed	18	3.93
			terrible	29	5.29	disaster	75	8.60			
						terrible	23	4.69			
						worst	18	4.05			
						awful	15	3.79			
Total	—		Total	66		Total	208		Total	18	

Positive											
Collocate	FaC	T-score	Collocate	FaC	T-score	Collocate	FaC	T-score	Collocate	FaC	T-score
—	—	—	—	—	—	best	53	5.52	—	—	—

The collocational profile of *waiting to happen*, however, contains much stronger evidence of negative semantic prosody. As Table 8.3 shows, twenty-three (46%) of the top-fifty collocates are negative at N-1 alone. Equally important is the fact that the negative collocates at N-1 combine for a total of 224 occurrences, or 54.5% of the 411 total occurrences.

It is worth noting yet again that eighteen of the twenty-six negative collocates at N-1 in Table 8.3 have t-scores lower than 2.0. and are, therefore, not considered statistically significant.

Table 8.3 Evaluative collocates in the 4:0 span of the T-Picture for *waiting to happen* (411) in the BoE

Negative												
N-4			N-3			N-2			N-1			
Collocate	Fa C	T-score	Collocate	Fa C	T-score	Collocate	Fa C	T-score	Collocate	Fa C	T-score	
tragedy	4	1.99	disaster	2	1.40	accident	5	2.22	accident	75	8.65	
disasters	2	1.41	accident	2	1.40	disaster	4	1.99	disaster	71	8.42	
disaster	2	1.40	bad	2	1.36	accidents	2	1.41	accidents	10	3.16	
overloaded	1	0.99	war	2	1.29	tragic	2	1.40	crisis	8	2.81	
			minefield	1	0.99	incident	2	1.40	tragedy	7	2.64	
			unprotected	1	0.99	crime	2	1.38	catastrophe	6	2.44	
			watergate	1	0.99	vitriolic	1	0.99	explosion	6	2.44	
			rapist	1	0.99	lurk	1	0.99	disasters	5	2.23	
			brawl	1	0.99	pests	1	0.99	scandal	4	1.99	
			unresolved	1	0.99	bse	1	0.99	incident	4	1.99	
			feud	1	0.99	disasters	1	0.99	injury	3	1.71	
			spill	1	0.99	unwanted	1	0.99	trouble	3	1.70	
									catastrophes	2	1.41	
									sacking	2	1.41	
									backlash	2	1.41	
									breakdown	2	1.41	
									riot	2	1.40	
									nightmare	2	1.40	
									upset	2	1.40	
									attack	2	1.37	
									trollop	1	1.00	
									tragedies	1	0.99	
									ambush	1	0.99	
									brawl	1	0.99	
									debacle	1	0.99	
									chernobyl	1	0.99	
Total	9		Total	16		Total	23		Total	224		

Positive												
efficient	2	1.40	heavens	1	0.99	lovable	1	0.99	saviour	2	1.41	
optimist	1	0.99	credible	1	0.99				champion	2	1.39	
momentous	1	0.99							winner	2	1.38	
Total	4		Total	2		Total	1		Total	6		

If statistical significance of collocates were a requirement for semantic prosody, evidence would be quite limited for the phrase *waiting to happen*. A similar trend is observable throughout profiles presented in this chapter.

In a somewhat surprising shift, collocational profiles for neither *accident* *waiting to happen* nor

disaster *waiting to happen* contain compelling evidence of semantic prosody. As Table 8.4 and Table 8.5 show, the numbers of evaluative collocates are once again quite low, and only one has an FaC greater than one (and none has a statistically significant t-score). Furthermore, only a few appear to be potentially directly syntactically connected to the node phrases.

Table 8.4 Evaluative collocates in the 4:0 span of the T-Picture for *accident waiting to happen* (75) in the BoE

Negative												
N-4			N-3			N-2			N-1			
Collocate	FaC	T-score	Collocate	FaC	T-score	Collocate	FaC	T-score	Collocate	FaC	T-score	
cliche	1	0.99	—	—	—	aids	1	0.99	vitriolic	1	0.99	
suicide	1	0.99							deadly	1	0.99	
injured	1	0.99							tragic	1	0.99	
problems	1	0.98										
Total	4		Total	—		Total	1		Total	3		
Positive												
efficient	2	1.41	heavens	1	0.99	lovable	1	0.99	—	—	—	
						classic	1	0.99				
Total	2		Total	1		Total	2		Total	—		

Table 8.5 Evaluative collocates in the 4:0 span of the T-Picture for *disaster waiting to happen* (71) in the BoE

Negative												
N-4			N-3			N-2			N-1			
Collocate	FaC	T-score	Collocate	FaC	T-score	Collocate	FaC	T-score	Collocate	FaC	T-score	
weaknesses	1	0.99	overloaded	1	0.99	spill	1	0.99	—	—	—	
			farce	1	0.99	blamed	1	0.99				
			outbreak	1	0.99							
			nightmare	1	0.99							
			tragedy	1	0.99							
			waste	1	0.99							
			warning	1	0.99							
Total	1		Total	7		Total	2		Total	—		
Positive												
right	1	0.99	—	—	—	—	—	—	—	—	—	
Total	1		Total	—		Total	—		Total	—		

Of course, it is expected to find that collocates of these final two iterations occur so infrequently for the very straightforward reason that there are far fewer instances of the phrase as it gets

longer. As Sinclair (2003, p. 125) notes, "we may observe that a two-word phrase, while not as rare as the arithmetical predictions would have it, is still not nearly as common as the words that make it up, and a three-word phrase is even less common."[1] Similarly, it is not at all surprising that the iterations of this phraseology have different collocates. The notion of "nested combinations" (Hoey, 2005) predicts as much. However, it is unexpected to observe that the numbers and frequencies of evaluative collocates varies so widely, given the fact that the 'core' element, *happen*, is so often claimed to have a negative semantic prosody (Sinclair, 1991, 2003; Bublitz, 1996; Partington, 2004, 2014).

Perhaps the clearest pattern emerging from the data presented thus far is the dominance of *accident* and *disaster* throughout. In Table 8.1, *accident* and *disaster* are the second and third most frequent negative collocates at N-3; in Table 8.2 they are the first and second most frequent collocates at N-2, and in Table 8.3 they are first and second at N-1. Notably, their frequencies in these positions decrease only very slightly from iteration to iteration, which appears to suggest that both *accident* and *disaster* are prominent collocates in these slots almost entirely by virtue of the fact that they are parts of the semi-preconstructed phrase *a|an accident/disaster waiting to happen*.

8.2.2 *Accident* and *disaster*: collocation or phraseology?

This section explores the notion that phraseology is 'responsible' for collocation and that corpus data examined so far calls into question whether it is accurate to consider *accident* and *disaster* 'collocates' of *happen*. As discussed above (Section 2.7), in this thesis I am "[focussing] on the generally accepted view of collocation as free combinations of a node and a collocate" (Barnbrook, Mason and Krishnamurthy, 2013, p. 164). One of the advantages of using the BoE

Picture software and PFTs is that they can provide indications of the degree of restrictedness characterizing node-collocate pairings. For example, as Table 8.1 has shown, *accident* is found only at N-3 in the 4:0 Picture for *happen*, illustrating that *accident* is likely not a "free collocate" of *happen*, but is instead highly restricted in its co-selection.

Table 8.6 illustrates the potential positional restrictions caused by phraseology; the FaC of *accident* and *disaster* will be greatly affected by the number of times the phraseology is found in the corpus.

Table 8.6 The 4:0 span of *happen* occupied by *a|an accident|disaster waiting to*

N-4	N-3	N-2	N-1	NODE
an	accident	waiting	to	happen
a	disaster	waiting	to	happen

This apparent positional restriction suggests quite strongly that *accident* is a 'collocate' in this position almost entirely by virtue of occurring in this semi-preconstructed phrase. Likewise, *disaster* is thirty-seventh at N-3 and does not occur anywhere else in the 4:0 Picture, indicating a similar degree of restrictedness. It is important to note that when instances of *a|an accident|disaster waiting to happen* are manually removed from the 43,759-line concordance of *happen* neither of these 'collocates' are found anywhere in the Frequency- or T-Pictures for the word form *happen*[2]. As Barnbrook et al. (2013, p. 165) argue, "these collocates are near the node because of the phraseology of the latter [the node]."

In contrast, the plural form, *accidents* appears at N-1 (twenty-ninth, FaC eighty-nine) and N-2 (fortieth, FaC eighty-seven), but not N-3 or N-4 of the T-Picture for *happen*, indicating a somewhat higher degree of collocational freedom. In fact, the phrase *accidents happen* occurs eighty-nine times in the BoE, and the top five collocates by frequency at N-1 are *most* (6), *when*

(4), *of* (4), *more* (4), *fatal* (4). That none of these collocates stands out as substantially more frequent suggests that there is no canonical variation (see Section 10.3 for detailed discussion), and further examination of the Pictures indicate that there are no salient larger iterations of the phraseology. Similarly, the phrase *accidents + [?] + happen* occurs 116 times. The top five collocates by frequency between *accidents* and *happen* are *will* (29), *do* (29), *can* (22), *that* (6), *could* (4). These suggest a colligational preference (Section 2.2) for modal auxiliary verbs, but no salient phrases are observed in the corpus data.

Conversely, the T-Picture for *CAUSE* reproduced in Table 6.5 above contains, for example, *problems*, *damage*, and *disease* in all eight columns of the 4:4 span, indicating that these collocations are indeed quite 'free' in that they collocate with CAUSE much more freely than *accident* does with *happen*.

8.2.3 Mode of evaluation

Focus now turns to the variable internal noun slot of the 'full' phrase *a|an NOUN waiting to happen*. In this section, it is argued that this semi-preconstructed phrase evaluates via the core meaning it acquires from the NOUN selected by the language user. The top ten noun collocates attested in the BoE of *a|an NOUN waiting to happen* are listed below in Table 8.7. The table shows the collocates, their raw frequencies, and the percentages of the 191 lines.

The collocates in Table 8.7 are very similar to those found in Table 8.3. However, the fact that they are now internal elements of the phrase means that, by definition, they are not evidence of its semantic prosody. It is perhaps worthy of note that this full iteration now satisfies the parameter of "semantic unity" (Gries, 2008, p. 6). As illustrated above, on the one hand, things that are *waiting to happen* do tend to be negative, as evinced in Table 8.3, but as the phrase

'grows' into the full, semantically unified form, corpus data shows the semantic prosody of *waiting to happen* 'switching off'.

Table 8.7 Top-ten noun collocates by frequency attested in the phrase *a|an NOUN waiting to happen* in the BoE

	NOUN	FaC	Percent
1	accident	66	34.6%
2	disaster	60	31.4%
3	tragedy	6	3.1%
4	catastrophe	4	2.1%
5	incident	4	2.1%
6	explosion	3	1.6%
7	injury	3	1.6%
8	sacking	2	1.0%
9	lawsuit	2	1.0%
10	breakdown	2	1.0%
11-44	—	39	20.4%
	Total	191	100.0%

Further research is required to determine whether semantic unity has a similar effect on the semantic prosody of other phrases [3]. However, as argued in the previous section, it is likely more correct to say that *waiting to happen* does not have a negative semantic prosody in the first place, but that *accident* and *disaster* appear in lists of collocates because of the phraseology. The full semi-preconstructed phrase *a|an NOUN waiting to happen* evaluates primarily via its evaluative core meaning in that it very rarely requires any additional context for its meaning to be observed in corpus data: "Speakers are aware of this [evaluative core] meaning, and the evidence from corpora confirms our intuition" (Mahlberg, 2005, p. 149).

All ten collocates in Table 8.7 are negative. The full profile reveals that positive evaluation is possible, but quite infrequent; *champion* (2), *winner* (2), *powerhouse* (1), *bargain* (1), *phenomenon* (1), and *success* (1) are all attested but they combine for only eight instances (4.2% of the 191 lines). Similarly, the full profile contains the neutral collocates *headline* (2), *hillsborough* (1), *takeover* (1), *novel* (1), *transfer* (1), *programme* (1), *idea* (1), and *deal* (1).

Together these represent only nine instances (4.7%). In all, then, the negative collocates account for more than 90% of the instances attested in the BoE.

Of this 90%, two collocates dominate the frequency list; *accident* (66) and *disaster* (60) together account for 66% of the 191 instances. For comparison, the enTenTen13 was also interrogated; Table 8.8 shows the top ten noun collocates of the 4,701 instances attested in the enTenTen13. The table again shows the collocates, their raw frequencies, and the percentage of the 4,701 lines.

Table 8.8 Top-ten noun collocates by frequency attested in the phrase *a|an NOUN waiting to happen* in the enTenTen13

	NOUN	FaC	Percent
1	accident	1,375	29.2%
2	disaster	1,350	28.7%
3	lawsuit	140	3.0%
4	tragedy	86	1.8%
5	catastrophe	83	1.8%
6	nightmare	74	1.6%
7	problem	67	1.4%
8	injury	65	1.4%
9	explosion	35	0.7%
10	adventure	34	0.7%
	[other]	1,392	29.6%
	Total	4,701	100.0%

The collocates and their relative frequencies in Table 8.7 and Table 8.8 are remarkably similar. Seven of the ten collocates are the same, although the order is slightly different. One notable variance, however, is that Table 8.8 contains one positive collocate, *adventure*. Most striking perhaps is the fact that *accident* and *disaster* are, again, the most frequent collocates. It is also worth noting that in both lists *accident* is the more frequent of the two, but only slightly. They differ by only six instances, or 3.2 percentage points the BoE. In the enTenTen13 they differ by only twenty-five occurrences, or only 0.5 percentage points.

The very small difference in frequencies between *accident* and *disaster* suggests a potential difficulty in selecting a canonical form of the phrase. The phrase *an accident waiting to happen* is, technically, the canonical form if frequency alone is the criterion by which such a form is selected. However, the data has shown that *disaster* is a very prominent alternative; it is the only other collocate even close to *accident* in frequency. It is so close that we must consider either that this semi-preconstructed phrase has two 'canonical' forms, or that the apparent "internal lexical variation" (Sinclair, 1991, p. 111) of the phrase is illusory. This would mean that that *an accident waiting to happen* and *a disaster waiting to happen* are not, in fact, variations of the same semi-preconstructed phrase, but are instead two very similar, but ultimately, discrete lexical items. That these phrases collocate so differently supports the latter argument, namely that they are two distinct lexical items. The potential importance of selecting a canonical form is discussed in greater detail in Chapter 10.

8.2.4 Qualitative Analysis

Recall that in the previous chapter it was demonstrated that the co-selection of the prepositional phrase *to bad people* reverses the core evaluative polarity of both *good things happen* and *bad things happen*. For this reason, this section discusses results of qualitative analyses of the sixty-six lines of *an accident waiting to happen* and the sixty lines of *a disaster waiting to happen* that were undertaken to determine whether context has an observable effect on the core evaluative meaning of the phrase revealed by the collocational profiles examined thus far.

8.2.4.1 Qualitative analysis of an accident waiting to happen

To begin, only three (4.5%) of the c. 200-character lines of *an accident waiting to happen* (66), shown below, contain any indication of positive evaluation at all; indications of positive

evaluation are underlined.

1. Yes, they're getting goals with style. <p> There's just no denying they're still **an accident waiting to happen** at the back. <p>
2. allow a decent recovery in large sections of the economy. Unless sterling, too, is `**an accident waiting to happen"** interest rates must fall.
3. Then there was Emergency Services Minister Merri Rose, thought by some to have been **an accident waiting to happen** but now starting to look like a potentially strong performer.

In lines 1 and 2 positive appraisals are found beyond sentence boundaries and do not appear to apply to the referent. In 3, the writer is contrasting his/her assessment of the referent against the opinions of *some*, but this does not reverse the core negativity of the phrase itself.

In nine lines (13.6%), the phraseology does give the superficial impression of evaluation via negative semantic prosody, meaning that the referent (i.e. that which is being characterized as *an accident waiting to happen*) has a negative core meaning. Four examples are presented below; negative referents are underlined.

4. gathering and harassment, black-bag jobs, and cover-up. These tactics made Watergate **an accident waiting to happen**. <p> Countering Chaos with CHAOS <p> and Domestic Surveillance <p> It is difficult to
5. shortcomings in the construction of the Chernobyl station, which she described as ` **an accident waiting to happen.'** The article was studiously ignored by the industry's authorities, although evidently
6. In retrospect I view Jim Jones as **an accident waiting to happen**. He could have been anywhere. Guyana was merely a backdrop for his Hollywood apocalypse
7. fact this sudden explosion of a real bullet from a theat-rical prop was **an accident waiting to happen**. For the real Philip Henslowe was, among other things, a churchwarden, and as such

Interestingly, three of these subjects (*Watergate*, *the Chernobyl station*, and *Jim Jone*s) are considered negative here because of cultural/historical associations, namely a significant political scandal, a nuclear disaster, and a mass suicide respectively. These would, originally,

have been considered evidence of textual evaluation, but as we have seen (Section 6.4.1), textual evaluation can become prosody (Mahlberg, 2005, p. 150). In addition, five of the subjects are phrasal, e.g. *this sudden explosion of a real bullet from a theatrical prop*, so would not appear on a standard collocational profile[4].

Most notably, fifty-four lines (83.3%) in the concordance contain neutral grammatical referents (almost always in the form of the grammatical subject of the copula verb immediately preceding the phraseology). In these lines, the negative core meaning of the phrase is being employed by the speaker to characterize as negative the otherwise neutral referents (underlined in the following examples):

8. <p> Later, I took the trainer aside and warned him that the girl was **an accident waiting to happen**: let her make up feeds or wash gear, I pleaded, but don't let her near young horses,
9. Western analysts have had the uneasy feeling that Russia was **an accident waiting to happen**. Well, it nearly did happen this fall when a group of political desperadoes,
10. The tax computations sent out to accountants and the self-employed are **an accident waiting to happen,"** says one taxman. `A vast number are being muddled up. The system is on the brink."

8.2.4.2 Qualitative analysis of a disaster waiting to happen

In the concordance of *a disaster waiting to happen* (60) there are seven lines (11.7%) containing general indications of positivity. In each case there are very clear lexical indications that the phrase is being used to characterize these positive items as unequivocally negative. Three of these lines are reproduced below; indications of positive evaluation are underlined:

11. THE Government's much vaunted private finance initiative has turned into **a disaster waiting to happen.**
12. at a level of passionate commitment never dreamed of by its pious founders: there is **a disaster waiting to happen.** Tennis and golf: the money gets bigger and better every year, but alas, the people
13. scientific advance in the field of nuclear waste treatment and help prevent what may be **a disaster waiting to happen."** The company admits it does not have the expertise to store the waste

For example, in 11 the contrast is accomplished by *turned into*; in 12 we see *alas*; and in 13 *admits* clearly signals the shift in evaluation.

In thirteen lines (21.6%) *a disaster waiting to happen* has a negative referent. Four examples follow; negative referents are underlined:

14. Britain's nuclear waste is **a disaster waiting to happen** </subh> <bl> By ROB EDWARDS </bl> <date> 19981219 </date> More than half of the 70 000
15. Some authorities are claiming that the new phylloxera outbreak was **a disaster waiting to happen**. Denis Boubals, Professor of Viticulture at Montpellier University, warned Californian
16. The fact is that Tuesday's nightmare was **a disaster waiting to happen** for a club divided from top to bottom. The pairing of mccann and Jock Brown has torn it
17. THE FIRE which caused the horrific deaths of the three Quinn children was **a disaster waiting to happen,** a near-inevitable consequence of unleashing the forces of anarchy and disorder on

Thirty-eight lines (63.3 %) have neutral grammatical subjects but also show clear evidence of negative evaluation elsewhere. Indications of negative evaluation are underlined in the following two examples:

18. the lambs and by the time I catch up with him there are three dead. I suppose this was **a disaster waiting to happen**. I am shaking with anger at my own stupidity.
19. Given the destruction caused by non-native species worldwide, some say there is **a disaster waiting to happen**. "We're playing Russian roulette with our salmon resources,"

In summary, no syntactic or contextual elements observable in the corpus data analysed appear to have the effect of reversing the polarity or changing the mode of evaluation of *an accident/disaster waiting to happen*. The majority of instances evaluate precisely as expected, namely via negative core meaning.

8.3 The ADJECTIVE Thing that can Happen

The semi-preconstructed phrase *the ADJECTIVE thing that can happen* was selected for the

current study because the phrase *the worst thing that can happen*, which appears to have a negative core meaning, is one of the MWUs uncovered using the methodology inspired by Danielsson (2007) (discussed in detail in Section 5.2.1 above). It is interesting to note that this is the only MWU in the list containing the general noun *thing* despite the high raw frequency of *thing* as a collocate of HAPPEN and its word forms (discussed at length in Chapter 6). Corpus data show that *the ADJ thing that can happen* appears to evaluate very differently from the other phrases containing *thing* discussed previously (see Chapter 7). The potential importance of *thing* as a phraseological element is discussed below in Section 8.3.4.3.

Corpus data presented in this section support arguments made above, namely that phraseological behaviour has a substantial observable effect on semantic prosody, and that seeking evidence of semantic prosody in collocational profiles appears to reveal the primacy of phraseology over collocation. There are some crucial differences in the profiles of the two phrases, however. Most notably, there are a considerable number of instances of *the ADJ thing that can happen* that evaluate via positive core meaning. This has pedagogical implications, which are discussed in detail in Chapter 10.

8.3.1 Collocational profiles of phraseological iterations of *the ADJ thing that can happen*

Discussion of the evaluative collocates in the 4:0 span for *happen* is found above in Section 8.2.1, but it may be worth repeating that corpus data shows very little evidence of semantic prosody in the 4:0 collocational profile of *happen* (see Table 8.1 above). Evidence for negative semantic prosody of *can happen* (2,299), however, appears to be much stronger. Table 8.9 below shows the evaluative collocates in the 4:0 span of *can happen*, their individual FaCs and t-scores. In addition, the table shows the total FaC for negative collocates at N-1.

As Table 8.9 shows, there are now eleven negative collocates at N-1, totalling fifty-nine occurrences, or 2.6% of the 2,299 instances of *can happen* in the BoE. Recall that there are only two negative collocates at N-1 in Table 8.1, *accidents* (89) and *mistakes* (18), which combine for a total of 107 (0.24%) of the 43,759 occurrences of *happen*. Therefore, although the negative collocates of *can happen* at N-1 are selected considerably more frequently than negative collocates at N-1 for *happen* alone, the raw percentage remains too low to support a strong argument for the negative semantic prosody of *can happen*.

Table 8.9 Evaluative collocates in the 4:0 span of the T-Picture for *can happen* (2,299) in the BoE

Negative												
N-4			N-3			N-2			N-1		
Collocate	FaC	T-score	Collocate	FaC	T-score	Collocate	FaC	T-score	Collocate	FaC	T-score
wrong	3	1.52	worst	80	8.92	worst	38	6.13	accidents	22	4.68
			worse	5	2.15	bad	7	2.48	mistakes	9	2.98
			terrible	4	1.94	terrible	5	2.19	bad	6	2.27
			warning	4	1.93	problems	5	1.98	accident	5	2.18
			bad	4	1.78	awful	4	1.96	disaster	3	1.67
									worst	3	1.63
									injury	3	1.62
									disasters	2	1.40
									rape	2	1.37
									stress	2	1.33
									worse	2	1.28
Total	3		Total	97		Total	59		Total	59	
Positive												
great	4	1.45	best	19	4.09	best	11	2.97	miracles	17	4.12
			good	8	2.15	wonderful	5	2.17	meaningful	2	1.39
			hope	5	2.05	miracles	4	1.99	special	2	1.02
						amazing	4	1.96			
Total	4		Total	32		Total	24		Total	21	

Evidence of semantic prosody is quite strong, however, in the profile for the three-word string *that can happen* (392), as illustrated in Table 8.10 below. Here we see more negative and positive collocates than in the previous profiles, but the prosody appears more strongly negative. There are thirteen negative collocates at N-1, sixteen at N-2, nine at N-1, and two at N-4.

Again, although the FaC values are, *prima facie*, much lower in Table 8.10, normalized percentages at N-1, for example, are much higher. Specifically, the fifty-two occurrences of negative collocates at N-1 now account for 13.3% of the 392 lines in the BoE concordance.

Table 8.10 Evaluative collocates in the 4:0 span of the T-Picture for *that can happen* (392) in the BoE

Negative												
	N-4			N-3			N-2			N-1		
Collocate	FaC	T-score	Collocate	FaC	T-score	Collocate	FaC	T-score	Collocate	FaC	T-score
lapses	1	0.99	accidents	2	1.41	worst	80	9.94	worst	37	6.07
breach	1	0.99	injury	2	1.39	worse	4	1.98	problems	3	1.67
			ridiculous	1	0.99	awful	3	1.72	accidents	2	1.41
			scandalous	1	0.99	terrible	3	1.72	imbalancing	1	1.00
			lunatic	1	0.99	bad	3	1.69	disintegrates	1	0.99
			sloppy	1	0.99	disgraceful	2	1.41	misbehaviour	1	0.99
			lapse	1	0.99	horrible	2	1.41	tangles	1	0.99
			terrifying	1	0.99	whacko	1	1.00	traumas	1	0.99
			debts	1	0.99	riskiest	1	0.99	pitfalls	1	0.99
						ugliest	1	0.99	tragedy	1	0.98
						nightmarish	1	0.99	risks	1	0.98
						damning	1	0.99	incident	1	0.98
						annoying	1	0.99	crisis	1	0.96
						stressful	1	0.99			
						outrageous	1	0.99			
						miserable	1	0.99			
Total	2		Total	11		Total	106		Total	52	
Positive												
imaginative	1	0.99	succeeds	1	0.99	best	19	4.31	best	8	2.75
exceptional	1	0.99	beneficial	1	0.99	pleasing	2	1.41	magic	3	1.72
						exciting	2	1.40	miracles	1	0.99
						greatest	2	1.39	bloom	1	0.99
						important	2	1.33	economical	1	0.99
						good	2	1.18	luck	1	0.98
						nicest	1	0.99			
						spontaneous	1	0.99			
Total	2		Total	2		Total	31		Total	15	

Complicating matters somewhat is the observation that for the string *thing that can happen* there are now nine negative collocates at N-1 (four fewer than in Table 8.10) with an FaC total of seventy-eight (twenty-six more than in Table 8.10), as shown in Table 8.11. These seventy-eight occurrences account for 61.4% of the 127 instances of the string in the BoE. From this

approach, then, Table 8.11 appears to contain by far the strongest evidence of negative semantic prosody of any of the iterations examined thus far, despite containing relatively fewer collocates.

Table 8.11 Evaluative collocates in the 4:0 span of the T-Picture for *thing that can happen* (127) in the BoE

Negative												
N-4			N-3			N-2			N-1			
Collocate	FaC	T-score	Collocate	FaC	T-score	Collocate	FaC	T-score	Collocate	FaC	T-score	
dying	1	0.99	—	—	—	lunatic	1	0.99	worst	66	8.12	
						worst	1	0.99	worse	3	1.72	
									disgraceful	2	1.41	
									horrible	2	1.00	
									whacko	1	0.99	
									damning	1	0.99	
									miserable	1	0.99	
									awful	1	0.97	
									bad	1	0.97	
Total			Total			Total			Total	78		
Positive												
bliss	1	0.99	benefit	1	0.98	—	—	—	best	19	4.34	
encouraging	1	0.99	important	1	0.96				greatest	2	0.99	
great	1	0.99							important	2	0.99	
									nicest	1	0.99	
									spontaneous	1	0.99	
									exciting	1	0.99	
									beautiful	1	0.99	
Total			Total			Total			Total			

Finally, Table 8.12 and Table 8.13 below show the evaluative collocates in the 4:0 span for the strings *worst thing that can happen* (66) and *best thing that can happen* (19). Neither table contains compelling evidence for semantic prosody even when the relatively low frequencies of the strings are accounted for. There are no evaluative collocates at either N-1 or N-2 in Table 8.12, and only one at N-3. At N-4 there are six negative collocates (12% of the fifty shown in the full BoE Picture), with a combined FaC of 9 (13.6% of the sixty-six occurrences of the phrase).

Like *accident* and *disaster* discussed earlier, the corpus data presented in this section contain

two dominant collocates. This time, however, one is negative, *worst*, and one positive, *best*. Again, an important observation is that at each stage of the analysis, these collocates are most frequent in the position they are found in the semi-preconstructed phrase *the ADJ thing that can happen*. For example, *worst* occurs eighty times at N-3 for *can happen*, eighty times at N-2 for *that can happen*, and 66 times at N-1 for *thing that can happen*; the same is true of *best*, the nineteen instances of which persist throughout the profiles of all three of these iterations.

Table 8.12 Evaluative collocates in the 4:0 span of the T-Picture for *worst thing that can happen* (66) in the BoE

Negative												
N-4			N-3			N-2			N-1			
Collocate	FaC	T-score	Collocate	FaC	T-score	Collocate	FaC	T-score	Collocate	FaC	T-score	
rape	4	1.99	dying	1	0.99	—	—	—	—	—	—	
vengeance	1	0.99										
unwilling	1	0.99										
agony	1	0.99										
consequence	1	0.99										
death	1	0.99										
Total	9		Total	1		Total			Total			
Positive												
—	—	—	encouraging	1	0.99	—	—	—	—	—	—	
Total			Total	1		Total			Total			

Table 8.13 Evaluative collocates in the 4:0 span of the T-Picture for *best thing that can happen* (19) in the BoE

Negative												
N-4			N-3			N-2			N-1			
Collocate	FaC	T-score	Collocate	FaC	T-score	Collocate	FaC	T-score	Collocate	FaC	T-score	
unhappy	1	0.99	—	—	—	—	—	—	—	—	—	
Total	1		Total			Total			Total			
Positive												
—	—	—	bliss	1	0.99	—	—	—	—	—	—	
			great	1	0.99							
Total			Total	2		Total			Total			

8.3.2 *Worst* and *best*: collocation or phraseology?

Corpus data in this section is presented to support the argument that, just as it is misleading to characterize *accident* and *disaster* as statistically significant collocates of *happen*, the same might be argued of *worst* and *best*.

The BoE contains ninety-one unique occurrences (0.20 per million words) of the 'full' iteration of *the ADJECTIVE thing that can happen* (two duplicate lines were removed from the ninety-three lines returned). As Table 8.14 illustrates, seven unique adjectives are attested. The enTenTen13, queried for comparison, contains 1,902 instances (0.10 per million words) with fifty-four distinct adjectives. Table 8.14 shows the top ten adjectives in the BoE and the enTenTen13 ranked by frequency, their raw FaC values, and the percentage of the total FaC.

Table 8.14 Adjectives ranked by FaC and percentages of total FaCs for *the ADJECTIVE thing that can happen*; 91 instances in the BoE and 1,902 instances in the enTenTen13

	BoE			enTenTen13		
	Collocate	FaC	% of total FaC	Collocate	FaC	% of total FaC.
1	worst	64	70.3%	worst	1,275	67.0%
2	best	18	19.8%	best	398	20.9%
3	only	4	4.4%	worse	70	3.7%
4	worse	2	2.2%	only	44	2.3%
5	greatest	1	1.1%	greatest	18	0.9%
6	other	1	1.1%	scariest	10	0.5%
7	nicest	1	1.1%	other	6	0.3%
8	—	—	—	same	6	0.3%
9	—	—	—	bad	4	0.2%
10	—	—	—	easiest	4	0.2%
				44 remaining	67	3.5%
	Total	91	100.0%	Total	1,902	100.0%

The BoE T-Picture for the word form *happen* ranks *worst* thirteenth at N-4 with an FaC of 155 and a t-score of 12.18 (cf. Table 8.1 above). As illustrated in Table 8.14, sixty-four of these 155 (41.3%) are found in the phrase *the worst thing that can happen*. Additionally, there are forty-

seven instances (30.3%) of the close variation, *the worst thing that could happen* in the BoE. Other more minor, but salient, variations include *one of the worst things that can happen* (13), and *one of the worst things that could happen* (2). These four variants together account for 126 (81.3%) of the 155 instances. Therefore, the *prima facie* evidence that *worst* is a statistically significant collocate of *happen* appears, in fact, to be a result of its presence in these few semi-preconstructed phrases[5]. When these 126 instances of the four major variants of the phrase are removed from the 43,759-line concordance of *happen*, the 'collocate' *worst* does not appear in either the BoE-generated frequency picture or T-Picture[6]. Likewise, *best* is ranked forty-first at N-4 of the T-Picture for *happen* with a t-score of 6.70 and FaC of 83. When *the best thing that can happen* (18) and *the best thing that could happen* (42) are removed, *best* disappears from the N-4 column of the BoE T-Picture and frequency-picture.

8.3.3 Mode of evaluation

This section discusses the observation that, like *a NOUN waiting to happen*, the phrase *the ADJ thing that can happen* does not evaluate via semantic prosody. Rather, it has a negative core meaning. It is noted, however, that when neutral collocates are selected (which is quite rare) the phrase is observed to have a negative semantic prosody. This is noted in support of the argument that the phrase appears to be primed to evaluate in toto, and, specifically, *thing* plays a significant role in the evaluative meaning of the phrase as a whole.

To begin, recall that in Table 8.14 an apparent tension has been revealed between the numbers of positive and negative collocates and their total FaCs. There are two negative collocates in the BoE list, *worst* (64) and *worse* (2) combining for sixty-six instances (72.5% of the ninety-one lines). In contrast, there are three positive collocates, *best* (18), *greatest* (1), and *nicest* (1), combining for only twenty instances (22%). A similar clash is found in the larger enTenTen13

data set, the top ten of which contains four negative collocates—worst (1,275), worse (70), scariest (10), and bad (4)—totalling 1,359 (71.5%) of the lines and three positive collocates—*best* (398), *greatest* (18), *easiest* (4)— accounting for 420 (22.1%) of the total.

This apparent clash is even more noticeable in the full list of collocates attested in the enTenTen13. Table 8.15 lists all fifty-four adjectives in the enTenTen13 concordance and their FaC values. The collocates are grouped by their core evaluative meanings, i.e. in three groups: negative, positive, and neutral.

Table 8.15 All adjectives in the phrase *the ADJECTIVE thing that can happen* in the enTenTen13 corpus; showing FaCs of negative, positive, and neutral adjectives

	Negative		Positive		Neutral	
	Collocate	FaC	Collocate	FaC	Collocate	FaC
1	worst	1275	best	398	only	44
2	worse	70	greatest	18	other	6
3	scariest	10	easiest	4	same	6
4	bad	4	nicest	3	last	4
5	worth[7]	2	sweetest	3	next	4
6	saddest	2	coolest	2	first	2
7	hardest	2	luckiest	2	weirdest	2
8	ugliest	2	rarest	2	second	2
9	riskiest	2	optimum	1	highest	2
10	stupidest	2	noblest	1	kinkiest	1
11	nastiest	1	kindest	1	biggest	1
12	startling	1	strongest	1	entire	1
13	darkest	1	finest	1	closest	1
14	severest	1	better	1	possible	1
15	extreme	1	simplest	1	strange	1
16	cruelest	1	next-best	1	smallest	1
17	sorst	1	healthiest	1	hottest	1
18	cruellest	1	ideal	1		
19	horriblest	1				
	Total	1,380		442		80

Table 8.15 shows that negative collocates outnumber positive by only one, at a ratio of 19:18, and the number of neutral collocates is also very close with seventeen. However, the nineteen negative collocates combine for an FaC total of 1,380, or 72.6% of the 1,902 lines. This total is

more than three times greater than the positive FaC total of 442 (23.2%), and more than seventeen times greater than the eighty instances (4.2% of the total) of the neutral collocates. In Chapter 6, similar clashes were discussed. Therefore, it was argued that, rather than the number of evaluative collocates, it is more appropriate to focus on which polarity is chosen when the initial pragmatic decision to evaluate is made.

The lists of adjectives in Table 8.14 and Table 8.15 illustrate two important observations. First, it seems that the semi-preconstructed phrase *the ADJ thing that can happen* is itself primed to evaluate. Whether this evaluation is positive or negative appears to depend entirely on which adjective is selected. Supporting this suggestion is the fact that the collocates labelled 'neutral' in Table 8.15 represent only eighty (4.4%) of the 1,902 occurrences of the phraseology. Moreover, many of these are observed to evaluate positively or negatively in the context of their concordance lines. For example, the third most frequent adjective in the BoE (fourth in the ententen13), *only*, has a neutral core meaning. However, as the three lines of *the only thing that can happen* reproduced below illustrate, this iteration of the phrase appears to evaluate via negative semantic prosody, which is observed in the referents (underlined).

20. That is **the only thing that can happen** if you play that way. You will concede a goal.
21. **The only thing that can happen** there is that they see the patients less.
22. **The only thing that can happen** to a slice is, it will get worse.'

These examples suggest a hypothesis that 4.2% of instances of *the ADJ thing that can happen* express neutral evaluation is a sizable overestimate and that truly neutral instances are exceedingly infrequent.

The collocational profiles above also suggest that in the vast majority of instances (1,822 or

95.8%) the adjective selected completes the phrase in a way that allows the evaluative meaning of the phrase to be observed in the absence of additional context. This is, as discussed above in Section 2.6, what Mahlberg (2005) calls "core meaning". In contrast, recall that in the previous chapter concordance lines were considered to evaluate via textual meaning "where our experience of language use and knowledge of the context is needed to interpret an example as evaluative" (Mahlberg, 2005, p. 150). The collocational profile of *the ADJ thing that can happen* suggests that no such experience or knowledge is required for more than 95% of instances.

8.3.4 Qualitative analysis of concordances

This section discusses analyses of two concordances comprising c. 200-character lines in the BoE. Despite collocational analysis indicating that *the ADJ thing that can happen* evaluates primarily via the core meaning of the adjective selected, qualitative analysis was undertaken to determine whether there is evidence in the concordance that the two major variants — *worst* (64) and *best* (18) — have their evaluative polarities reversed through linguistic or contextual factors not observable in collocational profiles.

8.3.4.1 Qualitative analysis of *the best thing that can happen*

In short, the qualitative examination of *the best/worst thing that can happen*, does not reveal a single phrase that reverses the evaluative polarity of this phrase, but there are instances in which the referents of *thing* warrant more careful examination. For example, six (33.3%) of the eighteen instances of *the best thing that can happen* in the BoE are marked in the sense that an unfavourable occurrence is expressed as having a favourable outcome. This is noteworthy because analysis of collocational profiles alone cannot account for marked instances like these:

23. songs burn out. In a while they die off, and that's **the best thing that can happen**. They can be rested; and then someone of another generation will discover... <p> R. Fisher: Yeah. Will find them, yeah.
24. **the best thing that can happen** to a sporting team is a major loss ... from defeat comes many more lessons than victory.
25. I don't wish a slip on anybody, I do want to acknowledge that sometimes it is **the best thing that can happen** to someone. Let me explain.

However, despite the apparent clash between negative referents and the positive core meaning of the phrase there is no evidence that evaluative polarity is reversed. For example, in line 23 the initial negativity of *burn out* and *die off* is reversed immediately by *they can be rested* and *another generation will discover*. Similarly, in line 24, *from defeat comes many more lessons than victory* justify the speaker's opinion that *a major loss* is *the best thing that can happen*. In 25, the speaker adds *let me explain*, and the extended context reveals a lengthy clarification.

The majority of the remaining twelve lines, are unmarked; they contain either explicitly positive or neutral referents that are explicitly characterized as positive. In the three examples reproduced below, positive referents are underlined.

26. She is a genuine, one-hundred-carat English creation, is Sue; for my money she is **the best thing that can happen** to anyone and, once committed, an enthusiastic, inventive and warm-hearted girl.
27. joining the stronger league, and ultimately acting as a feeder to the Brisbane Bears, will be **the best thing that can happen** to Aussie rules on the Coast.
28. this discreet gentrification is probably **the best thing that can happen** to what had become a sadly run-down corner of town.

In line 26, the referent *She* is neutral, but the 200-character line captures the explicitly positive characterization of *Sue* as *an enthusiastic, inventive, and warm-hearted girl*. In 27 and 28 *joining the stronger league* and *this discreet gentrification* are both positive and the extended context continues in the same vein.

8.3.4.2 Qualitative analysis of *the worst thing that can happen*

This section discusses the qualitative analysis of the sixty-four-line concordance of *the worst thing that can happen*, referred to briefly in Section 8.3.4.3 above. Of the sixty-four lines, five lines (7.8%) are marked in the sense that the writer appears to be contrasting the negative core meaning of *the worst thing that can happen* against a positive referent to create a surprising or ironic proposition. Positive referents are underlined:

29. <p> HE SAYS: `Celebrity is death. It's **the worst thing that can happen** to an actor."
30. **The worst thing that can happen** is that a young player has a lucrative contract and thinks, `That's it, I'm there," without achieving their potential."
31. In Buddhism the three terrible karmas are fame, beauty and fortune -- that's supposed to be **the worst thing that can happen** in your life."
32. **the worst thing that can happen** to a good neighbourhood restaurant is to be discovered. Too much favourable publicity often means farewell to reasonable prices, warm welcomes, careful cooking and unpretentious charm.
33. Everybody has made a wish at some time in his or her life. Sometimes **the worst thing that can happen** is that our wishes come true. This essay by CAROL GABEL involves such a wish.

Corpus data also shows that in twenty-two of the sixty-four lines (34.4%) *the worst thing that can happen* appears to be used with the intention of diminishing the core negativity (underlined) of the referent:

34. So try to keep it in proportion: an interview is not a firing squad **the worst thing that can happen** is that you don't get the job.
35. Face reality: a lot of people smoke, some get caught, it's not **the worst thing that can happen**. It's a set-back, but it's not murder, okay?"
36. Then trust it. Let it work. <p> After all, **the worst thing that can happen** is that you will miss the putt.

In Line 34 the phrase *try to keep it in proportion* sets up the expectation that *the worst thing* is either not meant to be interpreted literally, or is not as bad as might be expected. In line 35, the attempt to diminish the perceived negativity is clarified by, *it's a setback, but it's not murder, okay?* And in 36, the diminished negativity is presaged by, *After all*.

Finally, thirty-seven of the sixty-four instances (57.8%) of *the worst thing that can happen* are unambiguously negative. Two examples follow; negative referents are underlined:

37. I learned that being a mother means being able to forgive **the worst thing that can happen** to a mother: losing her child.
38. But we lost our love somehow. It's **the worst thing that can happen** to a person." She pulled a tissue from her handbag and wiped her eyes.

In these lines, it appears the referent is being characterized as literally the worst thing that can happen, and there are no indications that the speaker is intending to diminish the negativity in any way.

8.3.4.3 The general noun *thing* as 'evaluation carrier'

This section discusses the role of the apparently neutral general noun *thing* in the evaluative nature of this phraseology, i.e. two similar phraseological iterations — one containing *thing* and one without *thing* — appear to evaluate quite differently. This analysis supports the argument made above that *the ADJ thing that can happen* is a semantically unified phraseological whole which is itself primed to evaluate.

Li (2015, p. 188), in her discussion of the relationship between evaluation and phraseology (specifically *the ADJ thing (about n. / that cl.) is/was*), notes:

> [e]ven though the adjective in each sequence may determine the evaluative nature (or the potential evaluative meaning) associated with the sequence, it is the entire sequence that exhibits an evaluative use. In other words, the evaluative nature lies not just in each adjective in the sequence, but in each entire multi-word sequence.

In a similar way corpus data shows that the evaluative polarity of *the ADJ thing that can happen* is determined by the adjective that is selected, most often *worst*, followed by *best*. Additionally, in many of the concordance lines in the BoE, the general noun *thing* functions as the "evaluation

carrier" which "links the evaluation to the evaluated entity, and provides a background for the elements that express evaluative meanings" (Mahlberg, 2005, p. 154). Table 8.16 below reproduces an example in Hunston and Francis (2000, p. 134), showing how *thing* can act as evaluation carrier.

The evaluative pattern that more closely matches *the ADJ thing that can happen* is found in Mahlberg (2005, p. 152). Table 8.17 reproduces Mahlberg's (2005, p. 152) Table 6.2, which adds the category "evaluative context". An example from the BoE concordance of *the ADJ thing that can happen* has been added for comparison.

Table 8.16 Reproduction of Hunston and Francis (2000, p. 134) Table 5.10b showing the general noun acting as evaluation carrier

	Evaluative Category	Evaluation Carrier		Evaluated Entity
the	ADJ	general noun	v-link	to-inf
The	most difficult	thing	is	to score a goal…

Table 8.17 Reproduction of Mahlberg's (2005:152) Table 6.2, showing *thing* as evaluation carrier; the second example has been added here for the purpose of comparison to *the ADJ thing that can happen*

	Evaluative Category	Evaluation Carrier	Evaluative Context		Evaluated Entity
the	ADJ	general noun	about n	v-link	clause or n
the	surprising	thing	about chess	is	that computers can play it so well
the	worst	thing	that can happen	is	that you don't get the job

As Table 8.16 and Table 8.17 demonstrate, though the adjective does appear to give the phrase its core evaluative meaning, in fact, as Li (2015, p. 188) notes, each element in the phrase plays an important role in the expression of evaluation: "it is the entire sequence that is associated with evaluation rather than just the adjective."

This phrase lends itself to an interesting comparative analysis because both *the ADJ thing that*

can happen (64) and *the ADJ that can happen* (34) are salient variations. We might hypothesize that if *thing* did not add to or otherwise alter the evaluative meaning of the phraseology, there ought to be little observable difference in the corpus between the two. However, qualitative examination of the thirty-four lines of *the worst that can happen* reveals that they do, in fact, evaluate quite differently. Table 8.18 shows the raw frequencies and percentages in the BoE for each of three ways the phrases are observed to evaluate. The first row shows statistics for the 'marked' instances, i.e. where the phrase is used to characterize an apparently positive referent as, in fact, negative. The second row shows the data for instances expressing 'diminished negativity', i.e. where the referent is negative, but the proposition as a whole is intended to assuage negative feelings. The third row shows the statistics for instance that are simply 'negative', i.e. where the speaker appears to characterize the referent as literally *the worst*.

As Table 8.18 shows, *the worst thing that can happen* is used to diminish negativity in a considerable percentage of usages (34.4%).

Table 8.18 A comparison of the methods of evaluation of *the worst thing that can happen* and *the worst that can happen*

	the worst **thing** that can happen		the worst that can happen	
	Freq	Percent	Freq	Percent
Marked	5	7.8%	0	0.0%
Diminishing Neg.	22	34.4%	26	76.5%
Negative	37	57.8%	8	23.5%
Total	64	100.0%	34	100.0%

But the percentage is not nearly as high as it is for *the worst that can happen* (76.5%). Two such examples of this diminished negativity follow below:

39. In this metaphor, politics is a department store or an airport, a place in which **the worst that can happen** is that you might find yourself heading off on the wrong escalator,
40. **The worst that can happen** if we are wrong, Beloff suggests, is some embarrassment: `But again, we can handle that.

Of course, this difference is also reflected in the percentages of purely negative usages of the phrases; *the worst that can happen* is used literally in only 23.5% of lines, compared to 57.8% of lines containing *the worst thing*. That *thing* has a directly observable effect on evaluation in the instances cited here is yet another indicator that phraseology is a key factor in establishing an item's mode and polarity of evaluation. Namely, similar phrases — one containing *thing* and one without — appear to evaluate differently.

8.4 Conclusion

In this chapter, two phrases and related iterations have been analysed. The implication of the analyses presented in this chapter is that collocations and collocational behaviour provide only superficial evidence of semantic prosody. First, corpus data shows that iterations — or nested combinations (Hoey, 2005, 2007) — of the two phrases examined here have strikingly different collocational profiles. That nested combinations collocate differently is, of course, unsurprising, but differences in numbers and frequencies of specifically evaluative collocates observed in corpus data in the current study is noteworthy. Recall that the T-Picture for the word form *happen*, unsurprisingly, contains very little evidence of negative semantic prosody. It was noted that the strongest evidence is found at N-4 where the negative FaC is 312, but this represents only 0.7% of the 43,759 instances of *happen* in the BoE. There is somewhat more evidence of negative semantic prosody in the profile for *to happen* where the negative FaC at N-2 (where it is highest) is 208, or 1.9% of the 10,938 instances of this iteration in the BoE. It is only when we reach *waiting to happen* that we see clear evidence of negative semantic prosody; the negative FaC of 224 at N-1 represents 54.5% of the 411 instances.

A similar process of analysis for *the* ADJECTIVE *thing that can happen* led to similar results.

The two-word phrase *can happen* shows more evidence for negative semantic prosody than the lone word form *happen*, but the negative FaC at N-1 of *can happen* is still only fifty-nine, or 2.6% of the 2,229 occurrences. Collocational evidence of semantic prosody for the three-word phrase *that can happen* is stronger still with a negative FaC of 106 at N-2, which is 27.0% of the 392 occurrences. Finally, the profile of *thing that can happen* is strongest, with a negative FaC of seventy-eight at N-1, or 61.4% of the 127 instances of the phrase in the BoE

Second, the profiles of the shorter iterations presented in this chapter suggest that applying the term 'collocate' (in the statistically relevant sense) to the most frequently co-selected words (e.g. *accident*, *disaster*, *worst*, *best*) can be potentially misleading. This supports Barnbrook *et al*. (2013, p. 165) who argue, "[collocation] is a side-effect of phraseology". These 'collocates', which at first appear to be statistically significant, are in fact present in the profiles almost entirely because they are items comprising 'full' phrases. If these phrases are treated as discrete, semantically unified lexical items, and are removed from the concordance of *happen*, the collocates lose their statistical significance.

Additionally, the profiles of the semantically unified phrases also show that they do not have negative semantic prosodies. Instead, they tend to evaluate via core meaning. Whether these core meanings are positive or negative depends on the adjective or noun selected. One notable exception is in the relatively infrequent cases when a neutral adjective is selected to complete *the ADJ thing that can happen*; instances examined here do appear to evaluate via negative semantic prosody, though further research is required to confirm[8].

Finally, in addition to evaluating via core meaning, corpus data shows that the ostensibly very closely relatedly phrases *the worst thing that can happen* and *the worst that can happen*, in fact,

evaluate quite differently. This difference is accounted for by the inclusion of the apparently neutral general noun *thing* as evaluation carrier.

Notes

[1] Specifically, the word form *happen* (43,759) is almost exactly four times more frequent than *to happen* (10,938); the arithmetic calculation (which works on the problematic assumption of random distribution of words in the corpus), predicts that *to* (11,218,716) *happen* (43,759) would occur about 1,090 times in the BoE, when in fact it occurs almost exactly ten times more frequently.

[2] This is, of course, quite obvious if we return to the profile of *waiting to happen*. As Table 8.3 shows, the most frequent collocate at N-1, *accident*, occurs seventy-five times (18.2% of the 411 occurrences of the string). The second-most-frequent collocate, *disaster*, occurs seventy-one times (17.3%). The next-most-frequent evaluative collocate is the plural *accidents*, occurring only ten times (2.4%).

[3] Preliminary studies of *aware of what was happening*, for example, support this argument. The string *aware of* appears to have a fairly strong negative prosody, while the full MWU does not.

[4] The final referent is AIDS, found in the clause *a there will be a performance of AIDS: An Accident Waiting To Happen*. No amount of internet searching was able to reveal the nature of this performance, however, so no further discussion follows.

[5] The remaining twenty-nine instances comprise infrequent phrases such as *the worst thing ever to happen* (1), *the worst is bound to happen* (1), *the worst is unlikely to happen* (1), *the worst was going to happen* (1), and so on.

[6] *Worst* is also found at N-3 ranked thirtieth with an FaC of 124 and a t-score of 10.84. Here, the majority of instances comprise the phrase *the worst that can happen* (34), *the worst that could happen* (38), *the worst thing to happen to* (6) and other variations *the very worst that can happen*. When these semi-preconstructed phrases are removed from the 43,687-line concordance for *happen*, *worst* is not found in the picture sorted by t-score or frequency.

[7] Likely a misspelling of *worst*. Also, see *sorst* ranked 17 in the table.

[8] Though there are only 80 neutral instances (4.2%) of the 1,902 instances representing in the enTenTen13 (see Table 8.15 above), they do contribute evidence supporting the argument that the phrase *in toto* is primed to evaluate. Though not discussed in this thesis, a similar argument could be made about *a|an NOUN waiting to happen*; corpus data does seem to suggest that the entire phraseology is primed to evaluate.

CHAPTER 9:
EFFECTS OF REGISTER AND GRAMMATICAL PATTERNING ON SEMANTIC PROSODY

9.1 Introduction

This chapter returns to discussion of the semantic prosody of CAUSE tagged as a verb in the BoE and its sub-corpora (shown in Table 4.2 above). In the first part of the chapter (Section 9.2) collocational data are compared across the BoE and five specific registers. These registers are represented by individual or combined sub-corpora of the BoE as shown in Table 4.3 above. As the table shows, the **news** register comprises seven newspaper corpora, both broadsheet and tabloid[1]; **books** comprises two collections of fiction and non-fiction writing, one British and one American; the **spoken** register comprises two corpora of naturally occurring transcribed spoken conversations, one of British speakers and one of American speakers; **newsci** is a corpus of the weekly science magazine The New Scientist; and finally, **usacad** comprises academic texts printed in the United States.

Table 9.1 Combined BoE sub-corpora comprising registers examined in this chapter; tokens expressed in millions of words

News	Tokens	Books	Tokens	Newsci	Tokens	Spoken	Tokens	Usacad	Tokens
sunnow	44.7	brbooks	43.3	newsci	7.8	brspok	2.0	usacad	6.3
guard	32.2	usbooks	32.4			usspok	20.0		
econ	15.7								
oznews	34.9								
usnews	10.0								
indy	28.0								
times	51.8								
Total	217.6		75.8		7.8		22.1		6.3

Discussion in this section focusses on the lemma CAUSE and returns to the notion of semantic prosody as a phenomenon observed in lists of single-word collocates of a single-word node. As discussed in Chapter 6, such an approach can be problematic, but as Bublitz (1996, p. 13) argues, "[w]ith *cause*, almost all collocates occur within a span of 4 words to the left and 4 to the right.

This makes it easy to survey and handle even large KWIC lists."

Corpus data show that there are observable differences in collocational evidence for the negative semantic prosody of CAUSE in the registers examined. However, throughout this thesis it has been emphasized that "semantic prosody is not always a simple arithmetical function of the number of positive or negative items present in the stretch of discourse" (Morley and Partington, 2009, p. 142). For example, qualitative analysis shows that some ostensibly neutral collocates of CAUSE are found to contribute to its negative semantic prosody, and data show that at least one of the registers examined has a stronger semantic preference for this type of collocate.

The second part of the chapter (Section 9.3), examines the effects of grammatical patterning and register on the semantic prosody of CAUSE. Grammatical patterning has been included in the discussion of register primarily because one of the principle arguments made in Hunston (2007, p. 252) is that, "a word which is used in a certain way in most contexts is not necessarily used in that way in all contexts." To support this argument, Hunston employs twelve lines of CAUSE from the **newsci** sub-corpus, which show, in all but one instance, no evidence of negative semantic prosody. Figure 9.1 below shows samples of the twelve lines used by Hunston to illustrate; the lines have been shortened to more easily display the patterns of CAUSE in each.

Further, she argues that her twelve-line concordance illustrates that "CAUSE implies something undesirable only when human beings, or at least animate beings, are clearly involved" (2007, p. 253). Relevant to the structuring of this chapter is the fact that the twelve **newsci** lines selected by Hunston include only two grammatical patterns of CAUSE, namely ***CAUSE by* n** and ***CAUSE* n to-inf**.

Figure 9.1: Twelve-line concordance from Hunston (2007, pp. 252–253); lines are abridged to show the two frequent complement patterns **CAUSE *by* n** and **CAUSE n to-inf**

1. **cause** a smell to become less strong
2. **causes** a displacement current to flow through this capacitor.
3. **caused** by dark matter particles?
4. **caused** by short-term variations in weather.
5. **caused** by a galaxy
6. **caused** by a tidal interaction
7. **caused** by convection while the door is open.
8. **caused** by metal objects
9. **causing** it to spin more slowly.
10. **causes** the centre to be centrifugally displaced
11. **causing** the African bees to destroy each other.
12. **causing** the spin of the Earth to slow

As noted by Louw and Chateau (2010, p. 759): "Unfortunately, Hunston's paper does not clearly indicate how the examples were selected, or whether they were merely a random sample." Section 9.3, therefore, investigates the semantic prosody of CAUSE in four grammatical patterns, the two in Hunston's sample — **CAUSE *by* n** and **CAUSE n to-inf —** and two that are also frequently associated with CAUSE — **CAUSE n** and **CAUSE n n**. These patterns are examined in both the whole BoE and the **newsci** sub-corpus. Corpus data show that the two patterns comprising Hunston's figure do exhibit evidence of prosodic smoothing, but evidence is much stronger for one of them. Additionally, in both cases the smoothing is observable in the BoE and **newsci** but is noticeably stronger in the latter.

9.2 Effects of Register on Collocational Evidence for the Semantic Prosody of CAUSE

This section discusses the effects of register on the semantic prosody of the lemma CAUSE. The negative semantic prosody of CAUSE has been shown to be quite strong in general corpora comprising a large number of registers (see Stubbs' (1995) analysis of CAUSE in the LOB and analysis of CAUSE in the BoE presented in this thesis, especially section 6.2 above). As Table 9.2 shows, though, there are significant differences in the frequencies of CAUSE among the sub-corpora of the BoE; the table shows the BoE sub-corpora ranked by normalized frequency of

CAUSE from most to least frequent.

Table 9.2 Frequencies of CAUSE in BoE sub-corpora, showing raw and normalized frequencies per million words of text

Subcorpus	Raw Freq.	Normalized Freq.
newsci	4,682	593.0 /mil
usacad	2,680	422.6 /mil
bbc	6,023	323.7 /mil
usbooks	10,010	230.8 /mil
brbooks	3,485	221.7 /mil
econ	7,187	221.6 /mil
brephem	1,012	218.1 /mil
guard	6,914	197.9 /mil
oznews	6,225	192.9 /mil
npr	4,184	188.2 /mil
indy	8,113	181.3 /mil
strathy	4,999	178.1 /mil
sunnow	1,733	173.3 /mil
times	8,979	173.1 /mil
usnews	2,707	170.0 /mil
brmags	7,469	169.2 /mil
wbe	1,419	147.1 /mil
usephem	396	112.9 /mil
brspok	1,520	75.7 /mil
usspok	93	46.0 /mil
Total	89,830	200.3 /mil

In Table 9.2 the broadly scientific and academic registers, **newsci** and **usacad**, have the highest normalized frequencies of CAUSE. The lowest frequencies by far are found in the two spoken corpora, **brspok** and **usspok**. Also notable is that the normalized frequencies of **brbooks** and **usbooks** (combined as **books** below) are very close. These frequency groupings immediately suggest that register affects how CAUSE is used in different communicative contexts.

9.2.1 The effect of register on evaluative collocates of CAUSE

Table 9.3 shows the number of negative, positive, and neutral/unknown (including grammatical/function words) collocates in the top-fifty T-Lists for the lemma CAUSE in the BoE and five registers. As the table shows, there are sizable differences in the numbers of negative

collocates, suggesting that the negative prosody of CAUSE is indeed stronger in some registers than in others. BoE and **news** have the highest numbers of negative collocates (twenty-six and thirty respectively). **Books** and **newsci** have similar numbers (twenty-five and twenty-four). The lowest numbers of negative collocates are found in **spoken** and **usacad** (sixteen and fourteen).

Table 9.3 Numbers of negative, positive, and neutral collocates in the top-fifty T-Lists in the BoE and five registers for the lemma CAUSE

	BoE	News	Books	Newsci	Spoken	Usacad
Negative	26	31	25	24	16	14
Positive	0	0	0	0	0	0
Neutral/Unknown	24	19	25	26	34	36
Total	50	50	50	50	50	50

Table 9.3 shows that **usacad** has just over half the negative collocates that the general BoE corpus has, which appears to indicate extensive prosodic "smoothing" in the academic register (*cf*. Louw and Chateau 2010). However, this smoothing is not apparent in the scientific register represented by **newsci**, which, in fact, has substantially more negative collocates than **usacad**.

It is also difficult to account for the strong evidence of prosodic smoothing in **spoken**, which is very close in number of negative collocates to **usacad**. Indeed, the **spoken** data appear to contradict the suggestion by Louw and Chateau (2010, p. 763) that when "the Firthian context of situation is incomplete, negative prosody is smoothed and becomes neutral. […] However, when the context of situation is close to that of **normal language**, the negative prosody is reactivated" (emphasis added). If a more complete context of situation does in fact "reactivate" the prosody, then it is no surprise to see high numbers of negative collocates in **news**, which is arguably written for the widest audience and, therefore, is likely to contain the most "normal language" — see Partington (1998, p. 13) and Mahlberg (2005, p. 42) — and **books**, the

contents of which are also aimed at an audience of general readers. But this hypothesis on its own fails to explain why there are relatively few negative collocates in **spoken**, which arguably contains the most "normal language" of all the registers examined here.

The answer appears to lie in the FaC data, shown in Table 9.4 below. The table shows that the percentage of times a negative collocate in the T-list is chosen by a speaker in the **spoken** corpus — represented in the table by the FaC as a percentage of the total occurrences of CAUSE in the corpus — is, in fact, comparable to the other corpora.

Table 9.4 Comparison of numbers of negative collocates in the T-Lists in the BoE and five registers

	BoE	News	Books	Newsci	Spoken	Usacad
Total Occurrences of CAUSE	89,830	40,448	17,197	4,682	1,613	2,680
Number of Negative Collocates in the T-list	26	31	25	24	16	14
Total FaC of Negative Collocates in the T-list	28,380	15,523	4,903	1,572	591	418
FaC as a Percentage of Total Occurrences	31.6%	38.4%	28.5%	33.6%	36.6%	15.6%

The table also shows that the FaC percentage for **books** (28.5%) is somewhat lower than the others, and that of **usacad** (15.6%) is much lower. This strengthens the hypothesis that academic writing smooths the prosody of CAUSE. However, it is again evident that scientific writing represented by **newsci**, does not seem to have a smoothing effect; the negative collocate FaC percentage is nearly as high as that of **spoken**.

Corpus data focussing on only noun collocates at N+1 (*cf*. discussion in Section 6.3.1) produce quite different results, however. Table 9.5 shows that **spoken** has thirty-seven negative noun collocates at N+1, which is comparable to **news**, **books**, and **newsci**. However, this more focussed interrogation of the corpora reveals the distinction between **spoken** and the other registers, namely **spoken**'s total negative FaC is 206, or 61.1% of the 337 total occurrences. This is a substantially higher percentage than the total FaCs of negative collocates in the other

registers and supports the argument that the negative semantic prosody of CAUSE is, in fact, activated in normal usage.

Table 9.5 Comparison of noun collocates at N+1 of the T-Pictures for CAUSE

	BoE	News	Books	Newsci	Spoken	Usacad
Total Occurrences of CAUSE followed by a Noun	21983	10666	3856	1217	337	584
Number of Neg. Coll. at N+1 of the T-picture	41	38	38	35	37	27
Total FaC of Neg. Coll. at N+1 of the T-picture	8860	4802	1273	452	206	135
FaC as a Percentage of Total Occurrences	40.3%	45.0%	33.0%	37.1%	61.1%	23.1%

Additionally, **usacad** has twenty-seven negative collocates, thirteen more than in Table 9.4. However, smoothing is evident not in the raw number of negative collocates, but, again in the FaC totals which show that these twenty-seven negative collocates account for only 23.1% of the total occurrences of CAUSE followed by a noun in **usacad**. These two examples appear to support the argument for smoothing in academic registers and "reactivation" of the prosody in registers comprising "normal language".

It is also worth noting that corpus data comparing the numbers of negative collocates and their total FaC values indicates that CAUSE collocates much less freely in **spoken**. This means that while it appears that speakers make the pragmatic choice to evaluate negatively more often in the spoken register than in the other registers examined, they actualize that choice by co-selecting from a smaller set of collocates. Specifically, Table 9.6 below shows that, although *problems* (shaded in the table) is the highest ranked collocate at N+1 in four of the six registers examined, it accounts for 22.0% of all occurrences of *CAUSE* followed by a noun in **spoken**, which is almost three times more than the next most frequent use of *problems*, 7.5% in **news**. Similarly, the second-most frequent collocate in **spoken** is *trouble* (shaded), which accounts for 8.9% of all collocates at N+1. This means that when a noun immediately follows CAUSE in **spoken**, the noun is either *problems* or *trouble* more than 30% of the time. By comparison, the

combined FaCs of the top ten negative noun collocates at N+1 of CAUSE in the **news** register account for only 28.1% of the total instances.

Table 9.6 Top ten noun collocates at N+1 in T-pictures for CAUSE in six registers

	BoE	FaC	%	News	FaC	%	Books	FaC	%
1	problems	1488	6.8%	problems	795	7.5%	problems	202	5.2%
2	trouble	646	2.9%	trouble	382	3.6%	cancer	149	3.9%
3	cancer	551	2.5%	death	296	2.8%	trouble	106	2.7%
4	damage	511	2.3%	havoc	270	2.5%	damage	76	2.0%
5	concern	448	2.0%	concern	262	2.5%	pain	62	1.6%
6	death	432	2.0%	chaos	240	2.3%	people	51	1.3%
7	havoc	362	1.6%	damage	210	2.0%	death	43	1.1%
8	chaos	317	1.4%	offence	188	1.8%	aids	31	0.8%
9	aids	309	1.4%	cancer	169	1.6%	concern	31	0.8%
10	offence	238	1.1%	outrage	163	1.5%	irritation	29	0.8%

	Newsci	FaC	%	Spoken	FaC	%	Usacad	FaC	%
1	cancer	67	5.5%	problems	74	22.0%	cancer	36	6.2%
2	aids	49	4.0%	trouble	30	8.9%	people	18	3.1%
3	disease	47	3.9%	inflation	9	2.7%	cells	12	2.1%
4	problems	46	3.8%	delays	8	2.4%	disease	9	1.5%
5	damage	37	3.0%	chaos	8	2.4%	subjects	9	1.5%
6	interference	14	1.2%	damage	8	2.4%	resentment	8	1.4%
7	inflammation	13	1.1%	difficulty	8	2.4%	problems	8	1.4%
8	bse	11	0.9%	motion	7	2.1%	harm	7	1.2%
9	pain	11	0.9%	havoc	6	1.8%	illness	7	1.2%
10	concern	11	0.9%	arguments	6	1.8%	agents	7	1.2%

In summary, collocational data suggest that register — specifically academic writing — is responsible for smoothing the semantic prosody of CAUSE, and while it does seem possible to argue that "normal language", i.e. the **spoken** register, "activates" the negative semantic prosody of CAUSE, data suggest that collocates are more restricted in this register.

9.2.2 Effects of register on semantic preferences of CAUSE

Corpus data show that there are some significant differences in the semantic preferences of CAUSE across the registers examined. Table 9.7 shows the number of negative collocates at N+1 for CAUSE immediately followed by a noun grouped into their semantic preferences (see also

Table 1.2). As The table shows, **spoken** shows no apparent restrictions in semantic preference[2]. The damage/loss category is quite low with only one exemplar in the top fifty, which is comparable to damage/loss in the other registers. The strongest preference for diseases/injuries is in **newsci**. Data suggest that this is primarily due to numerous names of diseases and other medical conditions found in the corpus. In addition to *cancer* and *aids*, which are prominent in all registers examined, noun collocates of CAUSE at N+1 in **newsci** also include *blindness*, *bse*, *cjd*, *death*, *diarrhoea*, *disease*, *diseases*, *harm*, *headaches*, *illness*, *infection*, *inflammation*, *jaundice*, *malaria*, *pain*, *symptoms*, *tb*, *tuberculosis*, and *tumours*.

Table 9.7 A comparison of numbers of negative collocates at N+1 for CAUSE followed by a noun in each semantic preference category by register

	BoE	News	Books	Newsci	Spoken	Usacad
Diseases/Injuries/Symptoms	9	8	15	21	9	10
Psychological/Emotional Distress	17	15	12	6	9	6
Social Disruptions/Disturbances	5	7	5	2	10	1
Damage/Loss	7	4	2	3	1	1
Complications/Hindrances	3	4	4	3	7	8
Total	41	38	38	35	36	26

The stronger preference for names of diseases and symptoms of medical conditions in **newsci**, could help explain why there is no evidence that the semantic prosody of CAUSE is smoothed in that corpus. That is, virtually any disease name or medical symptom can be the grammatical subject or object (direct or indirect) of CAUSE, and as a category it is truly representative of the open class of nouns; it is impossible to name every disease and the list of disease names continually grows longer. Therefore, further study into claims that scientific writing has a smoothing effect might benefit from considering **health science writing** a discrete register.

Perhaps more significant is the fact that this group of diseases and symptoms are not necessarily representative of the register 'science writing' but is likely merely a consequence of the fact

that medical science appears to be a frequent topic in The New Scientist. A broader selection of science writing would almost certainly relegate these collocates to positions of obscurity well beyond the top fifty lists discussed here. Further research is required to determine whether other negative collocates would take their places, or whether the negative semantic prosody of CAUSE is in fact smoothed in science writing.

The notion of 'field', i.e. register-specific subject matter, may also explain why CAUSE is most frequent in **newsci**, as shown above in Table 9.2. That is, it could be argued that CAUSE is frequent in **newsci** because the subject matter discussed so frequently, namely diseases and symptoms, are the types of things that are said to be *caused*. While near-synonyms of CAUSE are grammatically possible, few are appropriate. As Chateau and Louw (2010, p. 763) argue, "the verb 'cause' is difficult to replace". For instance, many near-synonyms are phrasal verbs — *lead to, result in, bring about, give rise to* — that may be avoided by writers or editors for stylistic reasons. Others express subtle semantic differences that may make them inappropriate substitutes for CAUSE. For example, *engender* appears to have a relatively balanced semantic prosody observable in both positive (*confidence*, *respect*, *loyalty*, *trust*, and *optimism*) and negative collocates (*confusion*, *bitterness*, *hatred*, *guilt*, and *resentment*); but these collocates express a semantic preference for "a particular feeling, atmosphere, or situation" (CCED 2001:508), and so *engender* would not be an appropriate substitute for most of the preferences in Table 9.7.

It is also important to note that ten neutral collocates at N+1 for CAUSE in **newsci** — *birth*, *heart*, *skin*, *brain*, *lung*, *stomach*, *side*, *blood*, *breast*, *cell* — virtually without exception act as attributive nouns in negative phrases expressing the preference for diseases/injuries: *birth defects*, *heart disease*, *skin cancer*, *brain damage*, etc. These ostensibly neutral collocates are

found in other registers but they are clearly concentrated in **newsci** (**books** has six such collocates, **news** has four, **usacad** has three, and **spoken** only one).

9.3 The Effects of Grammatical Patterning on Semantic Prosody in the BoE and Newsci

This section discusses results of quantitative and qualitative examinations of the four CAUSE patterns (noted in Section 9.1 above) in the BoE and the **newsci** sub-corpus. The first pattern, **CAUSE n** was selected because of its obvious salience. That is, as explored in the previous section, observation of the direct objects of CAUSE in corpora (i.e. noticing that the things that are caused tend to share a semantic element) is the very foundation of the claim that CAUSE has a negative semantic prosody. The second pattern **CAUSE n n** was selected because it has been noted to have an observable effect on the semantic preference of CAUSE (*cf.* Stubbs, 2001c, p. 66). The remaining two patterns **CAUSE *by* n** and **CAUSE n to-inf** were selected for this study because, as explained in detail above, they account for all of the lines in a sample concordance used by Hunston (2007) to demonstrate the problematic nature of semantic prosody.

9.3.1 *CAUSE* n in the BoE

This section reports on the investigation into a 371-line BoE concordance for the grammatical pattern ***CAUSE* n**. In the previous section, single words tagged as nouns immediately following CAUSE were examined, regardless of what other patterns the lines contained. For this investigation, a concordance was constructed using two BoE queries[3], one requesting lines where CAUSE is immediately followed by a noun, the other where CAUSE is followed by a determiner and then a noun. Admittedly, this is something of an oversimplification of the noun group structure, and further research is required to discern whether pre-modifying adjectives and adverbs have observable effects on the semantic prosody of CAUSE. A concordance of 1000

lines (500 from each interrogation) was carefully edited and 628 inappropriate instances were removed, i.e. instances of *CAUSE* as a noun that have been incorrectly tagged by the tagging software as a verb (e.g. line 13 below), lines containing the contracted form of *because* (e.g. line 14), and instances in which ***CAUSE* n** is embedded inside one of the larger patterns (e.g. 15 and 16) were all discarded[4].

13.to identify the triggers than the **causes**. The word 'Thatcher"
14.tell them, just keep some of it, **'cause** those guys asked me
15. He then stopped suddenly, **causing** a car behind him to brake
16. Conservative Party was trying to **cause** panic among South Africa's

In lines where CAUSE is followed by a determiner, that determiner has been placed in the node position, so as not to split salient data between both N+1 and N+2. Table 9.8 illustrates this adjustment.

Table 9.8 An illustration of the how the determiners in the **CAUSE n** concordance were conflated with the node slot in the quantitative analysis

N-4	N-3	N-2	N-1	NODE	N+1	N+2	N+3	N+4
risks	an	infection	can	cause	death	and	whoever	is
the	stuff	that	s	causing the	problem	is	only	about

Table 9.9 shows the top fifty noun collocates at N+1 of the PFT created from the 371-line concordance. Forty-four of the fifty are negative (bold), indicating that the pattern **CAUSE n** does not have a smoothing effect on the negative semantic prosody of CAUSE in the BoE. In addition, there only two positive collocates (*appreciates* (1) at N-3 and *won* (2) at N-2) in the entire PFT. However, the frequencies of the collocates in Table 9.9 show evidence of the long Zipfian tail of collocates discussed in section 6.2.4 in that those ranked twenty-ninth to fiftieth all have an FaC of two. This suggests that additional quantitative investigation of the full concordance could reveal smoothing effects not apparent in the PFT.

Table 9.9 N+1 collocates in the 371-line BoE concordance for **CAUSE n**

	Collocate	FaC		Collocate	FaC
1	**problems**	26	26	**loss**	3
2	**disease**	12	27	**furore**	3
3	**cancer**	11	28	**disturbance**	3
4	**damage**	10	29	eye	2
5	**concern**	9	30	**infertility**	2
6	**problem**	8	31	**stress**	2
7	**havoc**	6	32	**health**	2
8	**trouble**	6	33	**embarrassment**	2
9	**injury**	6	34	**pain**	2
10	**aids**	6	35	**anger**	2
11	stir	6	36	**infection**	2
12	brain	5	37	**symptoms**	2
13	**accident**	5	38	**interference**	2
14	**discomfort**	4	39	**delays**	2
15	heart	4	40	**illness**	2
16	**confusion**	4	41	**tuberculosis**	2
17	**mayhem**	4	42	**congestion**	2
18	**storm**	4	43	**pollution**	2
19	**shock**	3	44	**tension**	2
20	**chaos**	3	45	**noise**	2
21	**inflammation**	3	46	**coma**	2
22	**anxiety**	3	47	**offence**	2
23	**casualties**	3	48	**consternation**	2
24	**miscarriage**	3	49	**destruction**	2
25	fire	3	50	**explosion**	2

Table 9.10 shows the total numbers of negative and positive collocates at each position in the 4:4 span of the 371-line BoE concordance for **CAUSE n**. The table shows that 279 (75.2%) of the 371 collocates at N+1 are negative, thus confirming that the pattern does not smooth the prosody of CAUSE in this register. Further examination of the full concordance shows that 341 (91.9%) of the 371 lines examined have at least one negative collocate in the 4:4 span.

Table 9.10 Number of evaluative collocates and percentages of the 371-line BoE concordance in the 4:4 span for the pattern *CAUSE* n

	N-4	%	N-3	%	N-2	%	N-1	%	N+1	%	N+2	%	N+3	%	N+4	%
Negative	24	6.5%	24	6.5%	38	10.2%	14	3.8%	279	75.2%	49	13.2%	49	13.2%	24	6.5%
Positive	4	1.1%	9	2.4%	5	1.3%	1	0.3%	0	0.0%	2	0.5%	4	1.1%	5	1.3%

Results of the qualitative examination of the 371-line concordance, shown below in Table 9.11,

are comparable. As Table 9.11 illustrates, in 340 lines (91.7%) the proposition expressed by **CAUSE n** is negative.

Table 9.11 Results of qualitative analysis of the 371-line BoE concordance for the pattern **CAUSE n**

Polarity	Freq	%
Negative	340	91.7%
Positive	2	0.5%
Neutral/unknown	29	7.8%
Total	371	100.0%

Only one line containing a negative collocate in the 4:4 span was labelled neutral/unknown in the qualitative analysis. In the quantitative analysis, *depression* was assumed to express the preference for psychological/emotional distress, and as such was considered negative, but this is not the sense employed in line 17, where the proposition is meteorological:

17. The depression **causing** the snow is now moving back towards central

Table 9.11 also shows that only two instances (0.5%) appear to express positive propositions (18 and 19 below):

18.possible for the first time, **causing** a communications revolution that is
19.or for the past three years, **causing** discounts"-the gap between the shares

Finally, twenty-nine (7.8%) are neutral/unknown, as in the following examples:

20. behave a certain way, just as genes **cause** eye color.
21.causes the water flow, the volume differential **causes** the heat flow.

Qualitative analysis also reveals that four of the six neutral collocates at N+1 of the PFT are attributive nouns (see also Section 9.2.1 above). These nouns, with only one exception (*eye color*), either modify a negative noun at N+2 (e.g. *brain tumours*/*damage*; *heart disease*/*attack*; *health problems*; *eye irritation*) or together with another neutral noun at N+2 form a negative

noun phrase (*side effects*). Qualitative analysis of the final two neutral/unknown collocates at N+1 failed to reveal whether any of the six instances of *stir* are intended by speakers to evaluate positively or negatively; however, all three instances of *fire* are found in negative propositions.

9.3.1.1 *CAUSE* n in newsci

This section discusses the study of 931 instances of ***CAUSE*** **n** in **newsci**. The concordance was created from the edited results of the same two BoE queries employed above. As above, lines not precisely typifying the pattern were removed[5]. Table 9.12 shows that 35 of the noun collocates at N+1 of the 931-line concordance are negative (bold). This is considerably lower than the forty-four negative collocates at N+1 for the pattern in the BoE, shown in Table 9.9.

Table 9.12 N+1 collocates of the 931-line **newsci** concordance for **CAUSE n**

	Collocate	FaC		Collocate	FaC
1	**disease**	55	26	**death**	5
2	**cancer**	53	27	**havoc**	5
3	**aids**	43	28	**confusion**	5
4	**problems**	28	29	**uproar**	5
5	**damage**	23	30	phenomenon	5
6	heart	15	31	**explosion**	5
7	**problem**	15	32	**accident**	5
8	**inflammation**	13	33	cell	4
9	skin	10	34	kidney	4
10	**pain**	10	35	**infections**	4
11	birth	9	36	memory	4
12	brain	9	37	liver	4
13	**bse**	9	38	**decay**	4
14	**infection**	9	39	**irritation**	4
15	side	8	40	nerve	4
16	**illness**	8	41	limb	4
17	stomach	8	42	**diseases**	4
18	**interference**	8	43	**blindness**	4
19	**pollution**	8	44	**cjd**	4
20	**trouble**	7	45	**tuberculosis**	4
21	**droughts**	6	46	**concern**	4
22	lung	6	47	**symptoms**	4
23	**tb**	6	48	changes	4
24	**tumours**	6	49	**furore**	4
25	**malaria**	5	50	**harm**	3

However, this is unsurprising given the noticeable shift in the semantic preferences, which now favour the diseases/injuries group. Specifically, in addition to the specific names of diseases (e.g. *cancer*, *bse*, *tb*, *malaria*, *cjd*), we see many of the same neutral attributive noun collocates (*heart*, *skin*, *birth*, *brain*, etc,) observed in the collocational profile above. These attributive nouns almost without exception modify negative nouns or together with a neutral noun comprise a negative noun group. As such, their presence in this position indicates that the negative prosody is much stronger than the thirty-five negative single-word collocates at N+1 suggests. That is, if we recall that the **n** in the pattern ***CAUSE*** **n** stands for 'noun group', not simply a single, discrete noun following CAUSE, then we can take into account the fact that these ostensibly neutral nouns at N+1 are in fact pre-modifiers in noun groups that are virtually all negative.

Quantitative analysis of the entire concordance, summarized in Table 9.13, further reveals that there are in fact 632 negative collocates (67.9%) at N+1 of the concordance. Another 147 (14.6%) negative collocates are found at N+2. Further analysis reveals that, in total, 87.1% (811) of the 931 **newsci** lines (compared to 91.9% in the BoE) have negative collocates in the 4:4 span.

Table 9.13 Numbers of evaluative collocates and percentages of the 931-line **newsci** concordance in the 4:4 span for the pattern **CAUSE n**

	N-4	%	N-3	%	N-2	%	N-1	%	N+1	%	N+2	%	N+3	%	N+4	%
Negative	65	7.0%	66	7.1%	114	12.2%	42	4.5%	632	67.9%	148	15.9%	102	11.0%	92	9.9%
Positive	9	1.0%	10	1.1%	8	0.9%	0	0.0%	4	0.4%	3	0.3%	6	0.6%	15	1.6%

Qualitative analysis of a 466-line concordance (every other line of the full concordance) of the ***CAUSE*** **n** pattern in **newsci** reveals that 416 instances (89.1%) express negative propositions. Only two lines express positive propositions, and both are instances of *caused excitement*:

22. Wisard **caused** excitement because, against expectations, it it was able to recognise individual human faces after only 20 seconds training.
23. Robert Redfield, […], **caused** excitement earlier this year when he told the international AIDS conference in Amsterdam that the vaccine reduced the levels of HIV in the bloodstream of infected people

In summary, the **CAUSE n** pattern does not appear to have an observable effect on the semantic prosody of CAUSE in either the full BoE or **newsci**. This is unsurprising, of course, because arguments that CAUSE has a negative semantic prosody tend to be supported by observations of strong collocational evidence that the things that are caused are most often negative; and these collocates are often found in noun groups immediately following CAUSE (see Stubbs (1995) and BoE data presented above).

9.3.2 *CAUSE* n n in the BoE

This section discusses analysis of the ditransitive pattern ***CAUSE*** **n n** in the BoE. Seven BoE queries returned at total of 1,841 lines. After careful editing, the concordance contained 423 lines. Initial investigations showed that there are no evaluative collocates at N+1 of the PFT for the 423-line concordance, meaning that none of the direct objects are, in themselves, negative. Instead, the evaluation is focussed on the indirect objects at N+2, where, as Table 9.14 shows, thirty of the collocates are negative (bold), and only two, *amusement* and *happiness*, are positive.

Eleven of the most frequent nouns in Table 9.14, including all of the top five, are also found in Table 9.9 above (collocates at N+1 for ***CAUSE*** **n)**. Notably, negative collocates in Table 9.9 that are not found in Table 9.14 are *illness*, *cancer*, *aids*, *tuberculosis*, *miscarriage*, *infertility*, *coma*, *symptoms*, *inflammation*, and *infection*. The absence of these collocates appears to confirm the restriction in semantic preference noted by Klotz (1997), cited by Stubbs (2001c, p. 66) who writes "in a double-object construction an illness is not possible," and notes that "frequent

exponents are nouns for feelings." Therefore, *headaches* in Table 9.14, as found in the ditransitive pattern, likely does not refer to physical pain, but the metaphorical sense of 'difficulty' or 'worry'.

Table 9.14 Collocates at N+2 of the 423-line BoE concordance for the pattern ***CAUSE*** **n n** (negative collocates are bold)

	Collocate	FaC		Collocate	FaC
1	**problems**	129	26	moments	2
2	**pain**	49	27	**hardship**	2
3	**concern**	20	28	**sadness**	2
4	**trouble**	17	29	amusement	2
5	**discomfort**	10	30	**agony**	2
6	**grief**	10	31	personal	2
7	**offence**	9	32	government	2
8	**distress**	9	33	visitors	2
9	**anxiety**	8	34	woodhead	1
10	**embarrassment**	8	35	crew	1
11	**stress**	6	36	hussein	1
12	lots	4	37	greenspan	1
13	physical	4	38	helens	1
14	**confusion**	4	39	**grave**	1
15	**alarm**	4	40	stephen	1
16	**anguish**	4	41	mundy	1
17	**misery**	3	42	police	1
18	plenty	3	43	**nausea**	1
19	**headaches**	3	44	**disease**	1
20	**difficulty**	3	45	happiness	1
21	**difficulties**	3	46	self	1
22	president	2	47	**disgrace**	1
23	**heartache**	2	48	tooth	1
24	**harm**	2	49	**irritation**	1
25	**shame**	2	50	**trauma**	1

There are, however, two apparent exceptions — *nausea* and *disease* — that resist this interpretation. These counter-examples suggest that we might have to take Stubbs' use of the term "illness" quite literally and limit this restriction to specific names of diseases (*cancer*, *AIDS*, *tuberculosis*):

```
24.        So small are the bugs which cause us disease, that they cannot
25.A long trip which in the past would cause him nausea and dizziness no
```

Again, because of the long tail of low-frequency collocates — seventeen collocates in Table

9.14 occur only once — the whole concordance was further examined in detail. Results of quantitative analysis of the full sample concordance is shown in Table 9.15. As the table illustrates, 333 (78.7%) of the 423 lines contain negative collocates at N+2.

Table 9.15 Numbers of evaluative collocates and percentages of the 423-line BoE concordance in the 4:4 span for the pattern **CAUSE n n**

	N-4	%	N-3	%	N-2	%	N-1	%	N+1	%	N+2	%	N+3	%	N+4	%
Negative	19	4.5%	14	3.3%	25	5.9%	9	2.1%	0	0.0	333	78.7%	26	6.1%	44	10.4%
Positive	10	2.4%	5	1.2%	4	0.9%	2	0.5%	0	0.0	4	0.9%	6	1.4%	6	1.4%

Further quantitative analysis reveals that 387 lines (91.5%) have at least one negative collocate in the 4:4 span.

Even stronger evidence of negative semantic prosody was observed in qualitative examination of the 423-line concordance. Table 9.16 below summarizes results of the qualitative analysis The table shows that 418 (98.8%) of the 423 instances of the pattern express unambiguously negative propositions.

Table 9.16 Results of qualitative analysis of the 423-line BoE concordance for the pattern **CAUSE n n**

Polarity	Freq	%
Negative	418	98.8%
Positive	5	1.2%
Neutral/unknown	0	0.0%
Total	423	100.0%

Only one line required additional context for disambiguation; in line 26, *embarrassment* falls just outside the 80-character span examined. This is because it is modified with the three-word determiner *a lot of* and the adjective *political*:

```
26.income taxes, an issue which has caused the President a lot of political
   embarrassment
```

None of the lines were labelled neutral/unknown, and there are only five positive instances, as follows:

```
27.  angelic instead of demonic, and cause him happiness instead of distress
28. don't know what I'm doing, which causes me amusement most of the time.
29.     he wondered why the idea had caused him amusement.
30.                  This revelation caused me fits of mirth.
31.         The cries of the wounded caused us joy, and increased our thirst
```

Data presented in this section illustrate that not only does ***CAUSE* n n** not smooth the negative semantic prosody of CAUSE toward neutral evaluation, the pattern overwhelmingly expresses negative propositions in the BoE. Once again, however, this result is not entirely unexpected. Hunston (2007, p. 253) argues that the negative prosody of CAUSE is likely activated in contexts where that which is caused directly or indirectly affects human (or other animate) beings and their endeavours. Since the direct objects observed in this pattern almost always refer to human beings — for example, the top five are *him* (78), *them* (75), *us* (59), *you* (51), *me* (48) — we expect the negative prosody of CAUSE to be activated in the majority of instances.

9.3.2.1 CAUSE n n in newsci

The ditransitive pattern **CAUSE n n** is quite rare in the **newsci** sub-corpus. Seven separate queries (see Note 2 at the end of this chapter for details) were employed in the corpus interrogation. These queries returned a total of 528 lines, but after extensive careful editing (again, removing exemplars of other patterns and mis-tagged instances) a sample concordance of only thirteen lines remained for study.

As with ***CAUSE* n n** in the BoE, there is no indication of prosodic smoothing evident for this pattern in **newsci**. The concordance is too small to support any strong claims, but all thirteen exemplars of the pattern, shown in Figure 9.2 below, are clearly negative.

Figure 9.2 Thirteen lines of the pattern **CAUSE n n** in **newsci**

```
32.            experiments that would cause animals 'pain
33.                  experiments that cause vertebrates pain or distress,
34.        experiments themselves may cause the animals pain and stress
35.                Both subscriptions cause us continuous problems," '
36.so the addition of the V-chip will cause them little trouble.
37.                                He caused Kasparov more problems
38.              of natural selection caused Darwin much anguish),
39.         her temples (her 'bumps") caused Cramer the greatest concern.
40.calling Eisler a German would have caused him great distress,
41.interpretations of the written word caused them enough anxiety.)
42.behaviour in the last few years has caused us great problems,
43.           in normal circumstances causes him chronic bowel inflammation
44.        part of their body that is causing them pain. <p>
```

The negative general nouns *problems* and *trouble* occur alongside the specific "nouns for feelings" (Stubbs, 2001c, p. 66): *anguish*, *concern*, *stress*, and *distress*. Again, however, we see one apparent exception to the preference restriction discussed above; line 43 contains *chronic bowel inflammation*, despite Stubbs' (2001:66) claim that Klotz found "an illness is not possible" in the ditransitive construction.

9.3.3 *CAUSE by* n in the BoE

This section examines the pattern ***CAUSE by*** **n** in the BoE. This is the first of the two patterns found in the twelve-line concordance in Hunston (2007, pp. 252–253). It is interesting to note that *by* is the highest ranking collocate of CAUSE in the BoE at N+1 by both t-score and frequency, with an FaC of 15,283 — which accounts for 17.0% of the 89,830 instances of CAUSE in the corpus — and a t-score of 120.08. The next most frequent collocate is *a* with an FaC of 7,848 (8.7%). The high frequency of *by* in this position may account, in part, for the 'random' selection of lines in Hunston's concordance. That is, the more frequent the pattern, the more likely it is to be 'randomly' selected.

The analysis began with two 500-line concordances. Only one line was removed from each,

leaving 998 lines for examination. Again, the node was 'expanded' to comprise CAUSE *by* and CAUSE *by* **determiner**. This was done to isolate noun collocates at N+1, as Table 9.17 shows.

Table 9.17 Node changes in PFT construction for the pattern ***caused by* n**

N-4	N-3	N-2	N-1	NODE	N+1	N+2	N+3	N+4
and	susceptibility	to	infections	caused by	destruction	of	blood	cells
likelihood	the	explosion	was	caused by a	bomb	planted	in	a

Quantitative analysis of the PFT indicates that the majority of negative collocates are found at N-1 and N+1, with twenty-eight and twenty-four respectively, shown in Table 9.18 below.

Table 9.18 Collocates at N-1 and N+1 of the 998-line BoE concordance for **CAUSE *by* n** (negative collocates highlighted in bold)

	N-1	FaC	N+1	FaC
1	was	85	**virus**	16
2	is	80	**lack**	13
3	been	53	gulf	9
4	be	51	**death**	9
5	**damage**	46	**hurricane**	8
6	are	37	**war**	7
7	were	32	exposure	7
8	**problems**	25	**bacteria**	6
9	not	16	**stress**	6
10	often	8	**recession**	6
11	**disease**	8	combination	6
12	**losses**	8	man	5
13	probably	7	fire	5
14	**stress**	7	**accident**	5
15	usually	7	fungus	5
16	**loss**	6	fact	5
17	**suffering**	6	drop	5
18	**disruption**	5	drug	4
19	**congestion**	5	changes	4
20	that	4	people	4
21	**death**	4	drugs	4
22	**diseases**	4	**infection**	4
23	**crisis**	4	**problems**	4
24	**chaos**	4	**inability**	4
25	**deaths**	4	millennium	4

	N-1	FaC	N+1	FaC
26	part	4	government	4
27	s	4	change	4
28	**dislocation**	4	fall	4
29	**pain**	4	build	4
30	**embarrassment**	4	**negligence**	4
31	possibly	4	**failure**	4
32	directly	4	way	4
33	**infections**	3	**pollution**	3
34	**inconvenience**	3	train	3
35	warming	3	vitamin	3
36	commonly	3	water	3
37	**upheaval**	3	**contamination**	3
38	**symptoms**	3	light	3
39	largely	3	food	3
40	**devastation**	3	**drought**	3
41	**problem**	3	gas	3
42	actually	3	**anxiety**	3
43	**difficulties**	3	mr	3
44	**pollution**	3	**loss**	3
45	brain	3	**bombing**	3
46	being	3	**flood**	3
47	**distress**	3	**injury**	3
48	pressure	3	**breach**	3
49	**delays**	3	**delay**	3
50	**tragedy**	3	sun	3
	Total FaC	598	Total FaC	237
	Total Neg FaC	181	Total Neg FaC	126

Recall that Table 9.9 above showed that forty-four of the top fifty collocates at N+1 for the

pattern **CAUSE n** in the BoE are negative. In Table 9.18, only twenty-four of the top-fifty collocates are negative, which suggests that some smoothing of the negative semantic prosody is occurring with this pattern. However, even though the N-1 slot is not an element of the ***caused by* n** pattern, it is not entirely unsurprising that there are so many evaluative collocates in this position, because the verb in this pattern is in the passive voice in most instances. That is, that which is caused, the grammatical object, is now found in the position of grammatical subject, i.e. preceding the verb. This also means that the total FaC at N-1 of 598 is somewhat misleading because BE (*was*, *is*, *been*, *be*, *are*, and *were*) occupies six of the top seven slots at N-1, totalling 338 instances. Further research may benefit from isolating only noun collocates at N-1. For these reasons, the **CAUSE n** and **CAUSE *by* n** data does not necessarily allow for direct comparisons, although their negative collocate FaC comparisons (see below) also indicate that the pattern smooths the prosody of CAUSE.

Table 9.19 below shows the numbers of negative, positive, and neutral/function word collocates in the PFT. As the table shows, the remainder of the 4:4 span contains relatively few evaluative collocates, and the only positive collocates are found at N-3.

Table 9.19 Numbers of evaluative collocates in PFT (4:4 span) of the 998-line BoE concordance for the pattern ***caused by* n**

	N-4	N-3	N-2	N-1	N+1	N+2	N+3	N+4
Negative	5	6	10	27	24	6	4	1
Positive	0	3	0	0	0	0	0	0
Neutral/Function word	45	41	40	23	26	44	46	49

The twenty-seven and twenty-four negative collocates at N-1 at N+1 indicate a relatively strong negative prosody. However, the total FaC of the negative collocates at N-1 of the PFT is only 181, or 18.1% of the 998 occurrences. Similarly, the total negative FaC at N+1 is only 126, or 12.6% of the total. These FaCs suggest a much higher degree of prosodic smoothing than the

raw numbers of negative collocates in Table 9.19. As discussed in Chapter 6, the low FaC values seem to suggest that the pragmatic decision to evaluate negatively is made much less often than the decision to eschew evaluation altogether. The relatively high number of negative collocates in the PFT may indicate only that when the pragmatic decision to evaluate negatively is made, speakers tend to select from a limited set of words.

Quantitative analysis of the whole concordance, shown in Table 9.20, confirms the low percentages of negative collocates at N-1 (33.3%) and N+1 (26.7%).

Table 9.20 Numbers of evaluative collocates and percentages of the 998-line BoE concordance in the 4:4 span for the pattern **CAUSE *by* n**

	N-4	%	N-3	%	N-2	%	N-1	%	N+1	%	N+2	%	N+3	%	N+4	%
Negative	112	11.2%	107	10.7%	152	15.2%	332	33.3%	266	26.7%	86	8.6%	84	8.4%	71	7.1%
Positive	16	1.6%	21	2.1%	7	0.7%	5	0.5%	8	0.8%	7	0.7%	13	1.3%	18	1.8%

Returning again to the pattern ***CAUSE*** **n,** we recall that the negative collocates at N+1 have an FaC total of 188, or 50.7% of the 371 lines examined. Further, 279 of the 371 ***CAUSE*** **n** lines (75.2%) have a negative collocate at N+1. By comparison, then, Table 9.20 does appear to show extensive smoothing.

However, further quantitative examination of the full concordance reveals that, in all, 761 (76.3%) of the 998 lines do have at least one negative collocate in the 4:4 span. This difference illustrates the danger of relying too heavily on statistically relevant collocates, PFTs, etc, or, indeed of relying on any one method of examination. While the information provided by statistical collocates and PFTs is often interesting and is frequently indicative of trends observable in the concordance, it is generally prudent to employ other methods of analysis (i.e. quantitative analysis of the whole concordance and qualitative examination) to confirm these

trends (*cf.* Tognini-Bonelli (2001, p. 24) for discussion of the potentially problematic nature of collocational profiles in identification of an item's semantic prosody).

Results of the qualitative analysis of a concordance of 499 lines (every second line of the 998 lines examined above) are shown below in Table 9.21. The table shows the frequencies and percentages of negative, positive, and neutral/unknown propositions in the concordance. As Table 9.21 illustrates, qualitative analysis of the pattern reveals no strong smoothing effect on the semantic prosody of CAUSE.

Table 9.21 Results of qualitative analysis of the 499-line **newsci** concordance of **caused by n**

Polarity	Freq	%
Negative	436	87.4%
Positive	5	1.0%
Neutral/unknown	58	11.6%
Total	499	100.0%

The 87.4% negative instances of the pattern ***CAUSE by*** **n** shows a very strong negative semantic prosody. At least fifty of the 499 lines examined have no negative collocates in the 4:4 span but evaluate negatively nevertheless. For example, in line 45 below, the negative collocate *imbalances* is found at N-5, and in 46 *problem* is at N-6. In line 47, there are no negative collocates in the *ca.* eighty-character line, but the extended context reveals that the proposition focusses on an *allergy* (at N-17, in the phrase *the most common allergy in women* […]) *to nickel*.

45. hormonal imbalances and all the fluctuations **caused by** puberty
46. The biggest problem with buying wine is that **caused by** the method of
47.to nickel. It most frequently affects the ear, **caused by** earrings

In conclusion, despite collocational evidence to the contrary, the pattern ***CAUSE by*** **n** does not appear to smooth the semantic prosody of CAUSE in the BoE.

9.3.3.1 *CAUSE by* n in newsci

This section reports on the results of investigation into the pattern **CAUSE by n** in the **newsci** sub-corpus. It was noted above that *by* is the highest ranked collocate at N+1 of CAUSE in the BoE by both frequency and t-score. The same is true of CAUSE in **newsci**, in which *by* occurs 1,130 times at N+1 (24.1% of the 4,682 occurrences), with a t-score of 32.6.

The same two queries used in the BoE study returned a total of 673 lines from **newsci** (none were removed). Again, as Table 9.22 shows, negative collocates are found primarily in the 1:1 span — twenty-three at N-1 and sixteen at N+1 of the PFT.

Table 9.22 Collocates at N-1 and N+1 of the 673-line **newsci** concordance for ***caused by*** **n** (negative collocates highlighted in bold)

	N-1	FaC
1	is	91
2	be	62
3	was	36
4	**damage**	32
5	are	30
6	were	24
7	been	19
8	**disease**	13
9	warming	11
10	**problems**	11
11	not	10
12	**pollution**	9
13	**diseases**	8
14	probably	7
15	**cancer**	5
16	that	5
17	changes	5
18	those	4
19	usually	4
20	**loss**	4
21	vibrations	4
22	cooling	4
23	**errors**	4
24	**deaths**	4
25	**accidents**	3

N+1	FaC
greenhouse	9
defect	6
virus	6
death	6
toxins	5
failure	5
radiation	5
water	5
impact	5
lack	5
moon	5
parasites	4
wind	4
drought	4
chemicals	4
changes	4
heat	4
exposure	4
light	4
fungus	4
build	4
use	4
deficiency	4
expansion	4
gene	4

	N-1	FaC
26	lines	3
27	**disturbance**	3
28	cells	3
29	**injuries**	3
30	all	3
31	energy	3
32	**stress**	3
33	**stresses**	3
34	change	3
35	level	3
36	effects	2
37	**carnage**	2
38	**disaster**	2
39	**storms**	2
40	**cancers**	2
41	it	2
42	**blurring**	2
43	one	2
44	**illnesses**	2
45	pattern	2
46	**fever**	2
47	**deformities**	2
48	being	2
49	neutrinos	2
50	**delays**	2
	Total FaC	469
	Total Neg FaC	123

N+1	FaC
sleepiness	3
driver	3
cfcs	3
impacts	3
damage	3
iron	3
shortage	3
bacteria	3
strokes	3
rotation	3
collisions	3
refraction	3
genes	3
hiv	3
infection	3
movement	3
bacterium	3
release	3
presence	3
sun	3
interaction	3
combination	3
reflection	2
waves	2
el	2
Total FaC	190
Total Neg FaC	66

This is five fewer negative collocates at N-1 and eight fewer at N+1 for the same pattern in the BoE (see Table 9.18), which does seem to indicate that the register is having an effect on the negative prosody of CAUSE. The remaining positions in the PFT also appear to display the effects of smoothing. Table 9.23 shows that outside of the 1:1 span, there are very few indications of negative evaluation; the overwhelming majority of collocates in the PFT are either neutral or function words.

Table 9.23 Numbers of evaluative collocates in PFT (4:4 span) of the 673-line **newsci** concordance for the pattern ***caused by* n**

	N-4	N-3	N-2	N-1	N+1	N+2	N+3	N+4
Negative	9	4	7	23	16	3	1	0
Positive	0	0	0	0	0	0	0	0
Neutral or Function word	41	46	43	27	34	47	49	50

However, further quantitative study encompassing the low-frequency collocates in the entire concordance shows a much less dramatic effect; 70.6% (425) of the 673 lines have negative collocates in the 4:4 span. While this is still lower than 76.3% in the BoE discussed in the previous section, it does not support a compelling argument that the **newsci** register is having a noteworthy effect on the semantic prosody of CAUSE in the pattern **CAUSE by n**.

Similarly, qualitative analysis reveals only very limited smoothing. Table 9.24 shows that in a 500-line concordance of ***CAUSE by* n** from **newsci**, 360 (72.0%) were found to evaluate negatively and none were positive.

Table 9.24 Results of the qualitative analysis of the 500-line **newsci** concordance of **CAUSED by n**

	Freq	%
Negative	360	72.0%
Positive	0	0.0%
Neutral/unknown	140	28.0%
Total	500	100.0%

In summary, quantitative analyses of ***CAUSE by* n** presented in this section show a slight smoothing of the semantic prosody compared to the analysis in the BoE. There are fewer negative collocates at N-1 and N+1, and fewer lines of the full concordance contain negative collocates in the 4:4 span (76.3% in the BoE compared to 70.6% in **newsci**). However, the smoothing effect is much more prominent in the results of the qualitative analysis. In the BoE, 87.4% of the lines were found to be negative, while in **newsci** this percentage is reduced to 72.0%.

9.3.4 *CAUSE* n to-inf in the BoE

This section examines the grammatical pattern **CAUSE n to-inf**, which is the second pattern in the figure used by Hunston (2007), and the final pattern examined in this chapter. A 1,483-line BoE concordance[6] was created from three carefully edited 500-line concordances. Once again, a brief note on the methodology used to create the PFT for this quantitative analysis is necessary. First, the node includes the determiner where applicable. This was done, as before, to ensure that salient noun collocates were not spread over two positions. This method also allows for the N+2 slot to be occupied entirely by *to*. As such, the range examined in the PFT has been extended to 4:5 to accommodate, as Table 9.25 illustrates.

Table 9.25 Illustration of the NODE and span changes in PFT construction for the pattern **CAUSE n to-inf**

N-4	N-3	N-2	N-1	NODE	N+1	N+2	N+3	N+4	N+5
which	in	turn	will	cause	imports	to	become	more	expensive
acting	in	ways	which	cause the	protests	to	grow	more	shrill

Immediately apparent in the PFT is the small number of evaluative collocates. Table 9.26 shows all the negative collocates in the entire 4:5 span. As shown in the table, there are only five negative collocates and one positive (*good*, ranked forty-eighth at N-4) in the 4:0 span.

Similarly, there are only ten negative collocates in the 0:5 span, all of which are verbs at N+3. These negative verbs have a total FaC of only 153, or just 10.3% of the 1,483 instances examined.

Table 9.26 Negative collocates in the 4:5 span of the PFT for the pattern **CAUSE n to-inf** in the BoE

N-4		N-3		N-2		N-1	N+1	N+2		N+3		N+4	N+5
Coll	FaC	Coll	FaC	Coll	FaC	Coll	Coll	Coll	FaC	Coll	FaC	Coll	Coll
fear	4	poor	3	problems	4	—	—	to	1,483	lose	53	—	—
				disease	4					miss	20		
				fear	4					suffer	15		
										crash	14		
										break	11		
										die	11		
										collapse	9		
										flee	7		
										fail	7		
										abandon	6		
										Total	153		

Quantitative examination of the full concordance, results of which are shown below in Table 9.27, reveal very few evaluative collocates throughout the PFT.

Table 9.27 Numbers of Evaluative collocates in the 4:5 span of the 1483-line BoE concordance of **CAUSE n to-inf** (N+2 is occupied by *to*)

	N-4		N-3		N-2		N-1		N+1		N+2		N+3		N+4		N+5	
	Freq	%	Freq	%	Freq	%	Freq	%	Freq	%	Freq	%	Freq	%	Freq	%	Freq	%
Neg	80	5.4%	101	6.8%	131	8.8%	60	4.0%	43	2.9%	—	—	313	21.1%	83	5.6%	69	4.7%
Pos	32	2.2%	30	2.0%	26	1.8%	12	0.8%	9	0.6%	—	—	36	2.4%	15	1.0%	30	2.0%

The percentages in Table 9.27 appear to show that the pattern ***CAUSE*** **n to-inf** has a strong smoothing effect on the semantic prosody. The strongest evidence of negative semantic prosody is at N+3 where only 311 (21.0%) of the verbs evaluate negatively, as in the following examples:

48. Financial innovation has **caused** banks to suffer declines
49. an explosive decompression which could **cause** the plane to crash.
50. which in the end may **cause** them to murder in the

Only 44 (3.0%) of the nouns at N+1 are negative, illustrated by in the following examples:

51. nerves trapped in the neck may **cause** pain to radiate down into the wrist
52. in the talks so far has **caused** frustration to grow on both sides
53. they tend to stick, **causing** strain to build up until

In all, only 673 lines (45.4% of the 1,483-line concordance) contain at least one negative collocate in the 4:5 span, indicating considerable smoothing of the negative semantic prosody of CAUSE.

For the qualitative analysis, a 495-line concordance (every third line of the 1483-lines analysed above) was examined. Results are summarized in Table 9.28. The table shows more negative propositions than the collocational profile suggests (recall that only 21.1% of verbs at N+3 are negative and there are very few negative collocates in the rest of the profile). However, 52.5% negative is still much lower than both the *ca.* 80% noted by Stubbs (1995, p. 4), and the percentages for the other patterns examined in this chapter which range from 72% to 98.8%.

Table 9.28 Results of qualitative analysis of a 495-line BoE concordance for the pattern **CAUSE n to-inf**

	Freq	%
Negative	260	52.5%
Positive	35	7.1%
Unknown	200	40.4%
Total	495	100.0%

The concordance also contains one interesting instance of what Louw (1993) calls "collocative clash" (see Section 1.2):

54. his assertion that a Yes vote would **cause** peace to break out everywhere,

Stubbs (1995, p. 2) writes: "Sinclair (ed. 1990:xi) […] points out that it is bad things which BREAK OUT." Stubbs corroborates from his own data citing the examples *violence*, *riots*, *war*,

etc. The collocate *peace*, then, 'clashes' with both *cause* and *to break out* in 54.

9.3.4.1 *CAUSE* n to-inf in newsci

The same three queries used above returned a total of 356 lines from **newsci**. Table 9.29 shows that there are only seven negative collocates in the 4:0 span of the PFT created from the **newsci** concordance, which is two more than in the same span for **CAUSE n to-inf** in the BoE (see Table 9.26). In the 0:5 span, there are nine negative verbs at N+3 (one fewer than in the BoE). At N+1 and N+4 in the BoE (see Table 9.26), there are no negative collocates, but the **newsci** PFT shows two at N+1 and three at N+4.

However, the total FaCs as a percentage of the total lines are almost identical. The total FaC of negative collocates at N+3 of Table 9.29 is thirty-nine, or 11.0% of the 365 lines. Recall that the total FaC of the negative collocates N+3 for the BoE was 153, or 10.3% of the 1,483 lines.

Table 9.29 Negative collocates in the 4:5 span for the pattern **CAUSE n to-inf** in **newsci** (N+2 contains *to* and is not shown)

N-4	N-3		N-2		N-1		N+1		N+3		N+4		N+5
	Collocate	FaC	Collocate	FaC	Collocate	FaC	Collocate	FaC	Collocate	FaC	Collocate	FaC	
—	fearful	1	irritant	2	collision	2	strain	2	collapse	9	wildly	2	—
			disease	2			galls	1	lose	6	murderers	1	
			parasite	1					deteriorate	5	prematurely	1	
			impurities	1					fail	4			
			instability	1					break	4			
									leak	3			
									explode	3			
									recoil	3			
									unravel	2			
—	Total	1	Total	7	Total	2	Total	3	Total	39	Total	4	—

Results of the quantitative analysis of the 4:5 span of the full 356-line concordance are shown below in Table 9.30. The 19.4% negative collocates at N+3 is very close to the 21.1% in the same position for ***CAUSE*** **n to-inf** in the BoE shown above in Table 9.27. In all, only 37.6%

(134) of the 356 lines contain a negative collocate in the 4:4 span.

Table 9.30 Numbers of Evaluative collocates in the 4:5 span of the 356-line **newsci** concordance of **CAUSE n to-inf** (N+2 contains *to*)

	N-4		N-3		N-2		N-1		N+1		N+2		N+3		N+4		N+5	
	Freq	%	Freq	%	Freq	%	Freq	%	Freq	%	Freq	%	Freq	%	Freq	%	Freq	%
Neg	14	3.9%	15	4.2%	23	6.5%	8	2.2%	9	2.5%	—	—	69	19.4%	16	4.5%	10	2.8%
Pos	3	0.8%	5	1.4%	2	0.6%	3	0.8%	1	0.3%	—	—	4	1.1%	6	1.7%	3	0.8%

Results of the qualitative analysis of the 356-line concordance are shown in Table 9.31. The table shows that the pattern expresses negative evaluation in 42.2% of the 356-line concordance. Only 3.9% are positive, and the remaining 53.9% are neutral/unknown.

Table 9.31 Results of qualitative analysis of a 356-line **newsci** concordance for the pattern **CAUSE n to-inf**

Polarity	Freq	%
Negative	150	42.2%
Postitive	14	3.9%
Neutral	192	53.9%
Total	356	100.0%

These results, compared to the 52.5% negative in the BoE lines in (Table 9.28) appear to show that **newsci** does have an additional smoothing effect.

9.4 Conclusion

9.4.1 Collocational evidence of prosodic smoothing by register

Corpus data show smoothing of the negative semantic prosody of CAUSE in collocational data from the written academic register represented by the **usacad** sub-corpus of the BoE. The collocational profile of CAUSE in the scientific register, represented by the **newsci** sub-corpus, somewhat surprisingly shows little indication of prosodic smoothing. The strongest prosody is observable in collocational data from **news** followed by **spoken**. The **news** register, it could be

argued, represents "normal language" (Louw and Chateau, 2010) and is thus precisely where, according to Louw and Chateau, we would expect the strongest indications of semantic prosody to be evident, and indeed, is where we find it. A significant observation is that although evidence for the semantic prosody of CAUSE is quite strong in the **spoken** register, the collocates used to activate the prosody appear to be somewhat restricted. That is, the T-List contains only sixteen negative collocates, but these collocates account for 36.6% of instances of CAUSE in the **spoken** register. By comparison, the T-List for **news** has thirty-one negative collocates, which account for 38.4% of CAUSE in that register.

It could be argued that collocational differences observed among these registers are partially explainable in Hallidayan terms of field, tenor, and mode. For example, **news**, in the form of newspaper reportage, is written (generally) for the widest audience of the five sets of sub-corpora interrogated and covers the widest range of topics to serve that audience. Partington (1998, p. 13) argues, "[newspapers] are the most widely read of long texts — almost everyone has considerable experience of newspaper language." Similarly, Mahlberg (2005, p. 42) writes "another way to view journalistic texts is to see them as representing mainstream English." **News**, therefore, could be argued to contain the highest concentration of normal written language.

While the mode (written) of **newsci** is essentially the same as **news**, both the field and tenor associated with *The New Scientist* differ in potentially important ways. *The New Scientist* is a weekly magazine targeting a more specific audience of both professionals and laypersons interested in scientific subject matter. Further, it may be something of an oversimplification to consider The New Scientist representative of the register of scientific writing. In fact, it is likely more accurate to place The New Scientist into a separate register of 'science reporting', as it

would seem to share characteristics with both journalistic and science writing.

Finally, **usacad** has the most specific field and tenor of the registers examined. The **usacad** sub-corpus comprises texts written for very specific audiences of professionals and students of explicitly academic (though not exclusively scientific) disciplines. In this register we might expect, therefore, to see the least 'normal' language, and, as a corollary, the least evidence of semantic prosody.

9.4.2 Evidence of smoothing as an effect of grammatical patterning

Table 9.32 shows a summary of the results of the quantitative and qualitative analyses of the four grammatical patterns associated with CAUSE in the BoE and **newsci**. The table shows that there is no evidence that the negative semantic prosody of CAUSE is smoothed toward neutral evaluation by either ***CAUSE* n** or ***CAUSE* n n**. However, qualitative analysis of a concordance of the pattern ***CAUSE by* n** revealed some smoothing in **newsci**, as only 72.0% of lines were labelled negative compared to 87.4% in the BoE.

Table 9.32 Summary of quantitative and qualitative results discussed in this chapter:

		% of lines with at least one neg coll in the 4:4 span	% of lines judged neg in qualitative analysis
CAUSE n	in BoE	91.9%	91.7%
	in newsci	87.1%	89.1%
CAUSE n n	in BoE	91.5%	98.8%
	in newsci	100.0%	100.0%
CAUSE by n	in BoE	76.3%	87.4%
	in newsci	70.6%	72.0%
CAUSE n to-inf	in BoE	45.4%	52.5%
	in newsci	37.6%	42.2%

This difference of fifteen percentage points is noteworthy, but perhaps not enough to support a persuasive argument that this pattern is responsible for smoothing the negative semantic prosody of CAUSE in scientific writing. Analysis of ***CAUSE* n to-inf** — both quantitative and qualitative — however, reveals compelling evidence for prosodic smoothing in both the general BoE corpus and **newsci**. The fact that both examinations show extensive smoothing in the general BoE corpus (Table 9.32 shows 45.4% and 52.5% negative respectively) indicates in that the pattern itself is responsible for the smoothing effect. Interestingly, however, the data also show that the effect is stronger in **newsci** suggesting that register, too, has a smoothing effect with respect to this pattern. It is perhaps no coincidence that the two patterns that appear to have a smoothing effect, **CAUSE *by* n** and **CAUSE n to-inf**, are the two patterns represented in the twelve-line concordance that Hunston (2007, p. 252) uses to illustrate the argument that "a word which is used in a certain way in most contexts is not necessarily used in that way in all contexts" (Hunston, 2007, p. 252). Corpus data show that register does have a smoothing effect but that the effects of these two patterns, especially **CAUSE n to-inf**, are much more pronounced.

Hunston (2007, p. 253) also argues that "[i]t seems reasonable to conclude that CAUSE implies something undesirable only when human beings, or at least animate beings, are clearly involved." However, corpus data in the form of a BoE Frequency Picture for the query cause@/VERB+NOUN+to+VERB (1,610), which is a simplified version of the **CAUSE n to-inf** pattern, shows that the nouns in the pattern are, in fact, often human. As shown in Table 9.33, six of the ten most frequent attested nouns explicitly refer to human beings — *people*, *investors*, *customers*, *men*, *women*, *children* — and the corpus reveals that one more, *others*, almost always refers to humans. Since this is the pattern that showed the highest degree of prosodic smoothing, it is somewhat surprising to see such a high degree of apparent human

involvement. Of course, as Table 9.32 above shows, the smoothing is not absolute (this pattern evinces negativity roughly half as often as the others), and so it is likely that Hunston is correct. These seven 'human' collocates account for only 285 (17.7%) of the 1,610 instances returned from this search. All of these and more could see CAUSE evaluate negatively and not have any effect on the smoothing observed.

Table 9.33 Top ten nouns by frequency in the BoE for the query cause@/VERB+ NOUN+to+VERB (nouns referring explicitly to people highlighted)

cause	NOUN	FaC	to-inf
	people	163	
	prices	42	
	investors	25	
	customers	24	
	others	23	
	cells	19	
	men	19	
	women	16	
	children	15	
	inflation	14	

Hunston (2007, p. 263) further writes that "rather than suggesting that register can make attitudinal meaning appear or disappear we might argue that particular registers select one lexical phenomenon more frequently than another." Corpus data confirms that the two patterns shown to smooth the negative semantic prosody of CAUSE are found more frequently in **newsci** and **usacad**. Table 9.34 shows the raw and normalized frequencies of the four CAUSE patterns in the BoE and five registers. These frequencies are of only one BoE query (concordances examined above were created from multiple queries), but they do indicate a notable difference. For example, as Table 9.34 shows, ***CAUSE* n to-inf** is found 20.0 times per million running words in the **usacad** sub-corpus, which is 5.5 times more than the 3.6 occurrences per million in the BoE. Similarly, the simplified version of ***CAUSE by* n** has a normalized frequency of 18.1 times in **usacad**, 1.8 times more frequent than the 10.1 occurrences per million in the BoE.

Table 9.34 Frequencies (raw and normalized) of the four patterns of CAUSE analyzed in this in the BoE and five registers

	cause n		cause n n		CAUSE by n		*CAUSE* n to-inf	
	cause@/VERB +NOUN		cause@/VERB +NOUN+NOUN		cause@/VERB +by+NOUN		cause@/VERB +NOUN+to+VERB	
	Freq	/mil	Freq	/mil	Freq	/mil	Freq	/mil
BoE	21,983	49.0	4,292	9.6	4,512	10.1	1,610	3.6
news	10,666	49.0	2,103	9.7	1,991	9.1	660	3.0
books	3,856	50.9	755	10.0	780	10.3	353	4.7
spoken	337	15.2	33	1.5	23	1.0	10	0.5
newsci	1,217	154.1	271	34.3	360	45.6	84	10.6
usacad	584	92.1	124	19.6	115	18.1	127	20.0

However, we have seen that CAUSE itself, regardless of pattern, is most frequent in these two registers (see Table 9.2 above and normalized frequency of the ***CAUSE* n** pattern in Table 9.34). It has been suggested that the high frequency in **newsci** is, at least in part, due to the subject matter often discussed in texts that comprise that specific sub-corpus, namely medical conditions, diseases, etc. Therefore, it is worthwhile looking at the more general version of the grammatical patterns — replacing CAUSE with any verb — to see if they are selected more frequently in the **newsci** and **usacad**.

Table 9.35 shows that these general patterns are more frequently selected in **newsci** and **usacad** than in the other registers. However, the **VERB *by* n** pattern is almost twice as frequent as **VERB n to-inf** in **newsci** and **usacad**.

Table 9.35 Frequencies (raw and normalized) for the patterns **VERB by n** and **VERB n to-inf** in the BoE and five registers

	VERB *by* n		VERB n to-inf	
	VERB+by+NOUN		VERB+NOUN+to+VERB	
	Freq	/mil	Freq	/mil
BoE	372,535	830.6	181,794	405.3
news	203,755	936.2	96,611	443.9
books	48,999	646.4	26,134	344.8
spoken	2,816	127.4	3,106	140.5
newsci	8,893	1,126.4	4,776	604.9
usacad	6,656	1,049.5	3,323	524.0

This higher frequency belies the fact that this pattern has been shown to have a small effect on the semantic prosody. The much larger effect has been observed for the less frequent pattern.

Finally, it is interesting to note how CAUSE ranks in the verbs selected in the patterns in Table 9.35. For the **VERB *by* n**, *caused* ranks consistently high in the Frequency Pictures in each register: eighth in the BoE, eleventh in **news**, fourth in **books**, fifteenth in **spoken**, second in **newsci**, and first in **usacad**. In contrast, none of the word forms of CAUSE are found in the top fifty of the Frequency Pictures of the BoE, **news**, or **spoken**, and in **books** *caused* is quite low at forty-sixth. In **newsci**, however, *cause* ranks much higher at seventeenth, and in **usacad**, *causes* is very highly ranked at sixth (and, unlike the other registers, *cause* and *caused* are also in the top twenty of the **usacad** picture).

Notes

[1] Further study might benefit from a more refined approach, i.e. dividing the **news** group into separate broadsheet and tabloid registers.

[2] The total numbers of collocates for **spoken** and **usacad** in Table 9.5 and Table 9.7 differ slightly because these two lists each contained one collocate that did not fit the preferences selected for this study. The **spoken** list contains *smogs* and **usacad** contains *erosion* which together suggest another preference (i.e. "detriments to the environment").

[3] All BoE queries employed in the investigation of the grammatical patterning of CAUSE:

***CAUSE* n**
cause@/VERB+NOUN+NOUN
cause@/VERB+PRON+NOUN
cause@/VERB+DET+NOUN+NOUN
cause@/VERB+ DET+NOUN+DET+NOUN

***CAUSE* n n**
cause@/VERB+NOUN+NOUN
cause@/VERB+PRON+NOUN
cause@/VERB+PRON+JJ+NOUN
cause@/VERB+NOUN+JJ+NOUN
cause@/VERB+JJ+NOUN+NOUN
cause@/VERB+DET+NOUN+NOUN
cause@/VERB+DET+NOUN+DET+NOUN

***CAUSE by* n**
cause@/VERB+by+NOUN
cause@/VERB by+DET+NOUN

***CAUSE* n to-inf**
cause@/VERB+NOUN+to+VERB
cause@/VERB+DET+NOUN+to+VERB
cause@/VERB+PRON+to+VERB

[4] In pilot studies, ***CAUSE*** **n prep** was treated as a discrete pattern, so these too were removed from the concordance. Preliminary analysis of this pattern showed no clear differences between it and ***CAUSE*** **n**, however, and so those results are not discussed in this thesis.

[5] The methodology did vary slightly here. I was initially concerned that the smaller **newsci** corpus would result in concordances too small to analyse effectively, so rather than start with 500 lines for each query, the lines for each query comprised the initial concordance: 1235 lines of cause@/VERB+NOUN were edited down to 712, and 728 lines of cause@/VERB+DET+NOUN were edited down to 219. My initial concern was obviously unfounded, and the final concordance analysed in this section is considerably larger than its counterpart in the previous section.

[6] This concordance is quite large relative to the others examined in this chapter. For each query, 500 lines or all lines (whichever was smaller) for each query were initially selected, and lines containing instances of CAUSE mis-tagged as a noun, adjacent duplicates, and embedded patterns were removed. Because the nature of the corpus tags and querying language make embedded patterns much less likely in these queries, i.e. the 'to-inf' acts as an easily identifiable (for the computer) pattern boundary, far fewer lines were required to be removed.

CHAPTER 10:
PEDAGOGICAL IMPLICATIONS AND APPLICATIONS

10.1 Introduction

This chapter discusses pedagogical implications and applications of the data presented in this thesis. For comparison to the BoE and enTenTen13 data, a 5.2-million-word corpus of Korean university student written English was compiled from six sources[1]. Table 10.1 below shows the token count for each corpus comprising the Combined Korean Learner Corpus (CKLC). The CKLC was queried using Antconc[2], and as in previous analyses, PFTs were created in Microsoft Excel.

Table 10.1 Corpora of Korean university students written English combined to form the Combined Korean Learner Corpus (CKLC) used in this thesis

Sub-corpus Name	Tokens[3]
CBNU Corpus of Written English	312,847
Chongshin University Corpus of Written English	190,606
The Gachon Learner Corpus[4]	2,606,008
ICNALE	136,358
NICKLE	941,012
YELC	1,099,473
Total	5,286,304

Section 10.2 discusses collocational evidence for the semantic prosody of CAUSE in the CKLC. Corpus data shows striking contrasts between the numbers of negative collocates observable in the Picture/PFT of CAUSE and their frequencies in the CKLC and the BoE. Section 10.3 discusses the potentially problematic nature of selecting and teaching a canonical form of the phrases examined in Chapters 7 and 8. Finally, Section 10.4 examines the pedagogical implications of the effects of grammatical patterning on the semantic prosody of CAUSE.

10.2 Semantic Prosody in the Learner Corpus

This section examines collocational evidence for semantic prosody in the CKLC and discusses

how such evidence can be helpful in the language classroom. In general, the more abstract aspects of meaning (see discussion of the lexical item in Section 2.2), like semantic preference and semantic prosody can be very helpful to L2 students because, as Hunston (2002, p. 20) has observed:

> […] native-speaker language teachers are often unable to say why a particular phrasing is to be preferred in a particular context to another, and the consequent rather lame rationale 'it just sounds better' is a source of frustration to learners.

Explicit reference to semantic prosody can help students to understand, for example, why certain expressions which are otherwise grammatically accurate seem awkward or even incorrect to a native speaker. For example, line 1 below, taken from the CKLC, is an example of this type of collocative clash. In this example the student evaluates that which is *provided*, namely *mini bar service*, negatively via the adjective *expensive*. However, as Stubbs (2001:65) has shown, *provide* has a positive semantic prosody. The T-List for PROVIDE in the BoE includes, for instance, *service*(*s*), *support*, *opportunit(y*/*ies*), *care*, *protection*, *assistance*, *help*, *benefits*, and *relief*. There are no negative collocates in the list.

1. but all hotels **provide** expensive mini bar service and room charge is very

The student's use of *provide* in line 1, then, creates a tension between the unconscious expectation of positive evaluation that the semantic prosody of *provide* creates for the reader and the negative evaluation of *expensive*. In the BoE there are 798 instances of *PROVIDE* ADJ *SERVICE* (where PROVIDE is tagged as a verb and SERVICE is tagged as a noun). While there are fourteen instances of PROVIDE *expensive* and thirty-four instances of *expensive* SERVICE, there are no occurrences of *PROVIDE expensive SERVICE* in the BoE.

However, as Section 10.2.1 shows, the negative prosody of CAUSE does not appear to be a

fundamental problem for the CKLC writers. Comparison of collocational profiles of CAUSE in the CKLC and the BoE reveal another pedagogically important characteristic of these L2 texts, namely they exhibit a much smaller active vocabulary than their L1 counterparts.

10.2.1 Collocational analysis of CAUSE in the CKLC

The lemma CAUSE occurs in the CKLC 3,445 times, or 651.68 times per million words, compared to 200.29 per million in the BoE. The apparent overuse of CAUSE in the CKLC could be because the corpus is not tagged for parts of speech, meaning that data cited here includes instances of CAUSE as both a noun and a verb, whereas the BoE data is of CAUSE tagged as a verb only. However, in the BoE, CAUSE is tagged as a verb 89,830 times and as a noun 36,952 times, a ratio of approximately 2.4 to 1. Applying the same ratio to the CKLC data results in a prediction of *ca.* 2,432 instances of CAUSE as a verb, or 460.06 instances per million words, which is still more than double the normalized frequency of CAUSE in the BoE.

It is also worth noting that, as Table 10.2 illustrates, the most frequent collocate of CAUSE at N+1 in the CKLC is not *by* (see Section 9.3.3 for discussion of CAUSE *by*). Instead, *of* tops the list with 406 occurrences (11.8% of all instances of CAUSE in the CKLC), suggesting that Korean university students use CAUSE as a noun more frequently than native speakers. Similarly, *the* is much more prominent in the CKLC List (4:4 span inclusive) than it is in the BoE List. As Table 10.3 below shows, *the* is the most frequent collocate in the CKLC T-List, which also suggests that Korean student writers tend to use CAUSE as a noun more frequently than L1 language users. Still, these frequency differences may not be enough to account for the apparent overuse of CAUSE in the CKLC[5].

The comparatively low number of negative collocates in the CKLC T-Lists shown in Table

10.3 below also stand out; only four of the top twenty-five CKLC collocates are negative, compared to fourteen in the BoE data (and only eight of the top fifty CKLC collocates are negative compared to twenty-six in the full BoE List).

Table 10.2 Top Twenty Collocates CAUSE ranked by frequency at N+1 and N+2 in the BoE and CKLC, negative collocates in bold

	BoE						CKLC					
	N+1	FaC	%	N+2	FaC	%	N+1	FaC	%	N+2	FaC	%
1	by	15,283	17.0	to	5,956	6.6	of	406	11.8	**accident**	147	4.3
2	the	9,004	10.0	the	5,413	6.0	by	300	8.7	to	140	4.1
3	a	7,848	8.7	a	2,825	3.1	a	225	6.5	the	132	3.8
4	**problems**	1,488	1.7	and	2,228	2.5	many	153	4.4	**problems**	89	2.6
5	an	1,471	1.6	in	2,168	2.4	the	148	4.3	accidents	87	2.5
6	him	1,250	1.4	**damage**	1,510	1.7	serious	75	2.2	and	82	2.4
7	it	1,183	1.3	of	1,397	1.6	an	64	1.9	traffic	76	2.2
8	some	1,169	1.3	**problems**	1,319	1.5	i	53	1.5	of	71	2.1
9	them	1,114	1.2	by	816	0.9	to	52	1.5	car	69	2.0
10	to	926	1.0	<p>	784	0.9	car	51	1.5	**problem**	53	1.5
11	more	899	1.0	for	707	0.8	is	49	1.4	i	43	1.2
12	you	773	0.9	an	509	0.6	**obesity**	47	1.4	effects	41	1.2
13	her	755	0.8	or	481	0.5	big	47	1.4	lot	38	1.1
14	any	649	0.7	much	471	0.5	that	44	1.3	a	37	1.1
15	**trouble**	646	0.7	but	469	0.5	people	41	1.2	**obesity**	34	1.0
16	serious	628	0.7	s	457	0.5	**accident**	40	1.2	people	33	1.0
17	me	614	0.7	among	452	0.5	traffic	38	1.1	in	31	0.9
18	us	567	0.6	**concern**	442	0.5	side	36	1.0	**diseases**	31	0.9
19	**cancer**	551	0.6	lot	414	0.5	more	29	0.8	so	29	0.8
20	**damage**	516	0.6	**problem**	391	0.4	it	28	0.8	**cancer**	28	0.8
21	such	511	0.6	**harm**	373	0.4	**accidents**	26	0.8	**disease**	26	0.8
22	this	496	0.6	stir	364	0.4	some	24	0.7	serious	25	0.7
23	in	480	0.5	**death**	353	0.4	so	22	0.6	effect	25	0.7
24	no	480	0.5	**pain**	349	0.4	other	22	0.6	s	24	0.7
25	**concern**	449	0.5	**trouble**	341	0.4	they	22	0.6	this	24	0.7

Additionally, the nature of the collocates is also quite different. There are at least thirty grammatical/function words in the full CKLC list, compared to only eighteen in the BoE, which suggests that the Korean student writers likely have a more limited active vocabulary than L1 English users. This, in itself, is not especially surprising, but comparison of the negative FaC values shows, below, that the negative prosody of CAUSE appears to be activated at levels comparable to those in the L1 data.

Table 10.3 Top twenty-five Collocates in the T-Lists for CAUSE in the BoE and CKLC, negative collocates bold

	BoE	FaC	T-score	CKLC	FaC	T-score
1	by	18,267	109.20	the	1,060	30.35
2	**problems**	4,838	67.05	of	973	30.06
3	**damage**	4,323	64.92	it	856	28.46
4	which	5,964	53.78	and	711	24.62
5	can	5,321	52.55	can	622	24.35
6	to	25,805	48.73	is	650	23.59
7	that	11,507	39.18	to	603	22.20
8	may	2,740	36.99	a	518	20.99
9	could	2,841	35.90	**accident**	411	20.21
10	**trouble**	1,408	35.66	that	433	19.47
11	**cancer**	1,366	35.34	by	352	18.47
12	**pain**	1,286	34.37	many	299	16.75
13	**disease**	1,278	34.16	in	323	16.03
14	**harm**	1,110	32.80	so	284	15.60
15	**death**	1,318	32.08	this	260	15.46
16	**concern**	1,106	31.43	i	365	15.12
17	has	4,402	29.63	**accidents**	224	14.93
18	**serious**	989	28.31	**problems**	208	14.35
19	**loss**	883	26.89	car	198	13.88
20	some	2,314	26.61	people	218	13.47
21	**problem**	1,029	26.00	traffic	178	13.24
22	**distress**	675	25.65	because	198	13.13
23	this	4,498	25.63	are	210	13.07
24	among	997	25.57	**obesity**	170	13.02
25	**injury**	764	25.57	driving	166	12.63

Table 10.4 and Table 10.5 show the negative collocates of CAUSE in the CKLC in the 4:0 span and 0:4 span respectively, ranked by frequency. There are no positive collocates at all in the entire 4:4 span.

Table 10.4 Negative collocates of CAUSE (3,445) in the 4:0 span in the CKLC

N-4		N-3		N-2		N-1	
Collocate	FaC	Collocate	FaC	Collocate	FaC	Collocate	FaC
dangerous	30	dangerous	27	accidents	23	accident	29
bad	15	violent	27	violent	22	problems	19
obesity	13	bad	15	accident	17	accidents	17
		obesity	13	bad	16	problem	14
				problem	15	obesity	14
				obesity	14	punishment	11
				punishment	13		
Total	58	Total	82	Total	120	Total	104

Table 10.5 Negative collocates of CAUSE (3,445) in the 0:4 span in the CKLC

N+1		N+2		N+3		N+4	
Collocate	FaC	Collocate	FaC	Collocate	FaC	Collocate	FaC
obesity	47	accident	147	accident	128	accident	33
accident	40	problems	89	accidents	44	problems	21
accidents	26	accidents	87	problems	40	accidents	17
bad	21	problem	53	problem	36	bad	12
terrible	21	obesity	34	disease	31		
cancer	20	diseases	31	obesity	23		
problems	19	cancer	28	stress	13		
dangerous	15	disease	26				
harm	14	damage	17				
disease	13	death	16				
death	12	dangerous	13				
harmful	12	crime	13				
		bad	12				
Total	260	Total	566	Total	315	Total	83

Compared to the numbers of negative collocates of CAUSE in the BoE, the numbers in the CKLC are notably small, but the FaC values as percentages of the total occurrences at each slot is often higher. As Table 10.6 below shows, for example, at N-4 of the BoE T-Picture, there are eighteen negative collocates with a total FaC of 1,112, which is 1.2% of the 89,830 occurrences of CAUSE in the BoE. Comparatively, there are only 3 negative collocates at N-4 in the CKLC, but they account for 1.7% of the 3,445 instances of CAUSE in the CKLC. As the table shows, the CKLC difference is +0.5 percentage points. The largest differences, 6.1 and 4.4 percentage points, are at N+2 and N+3 respectively.

Table 10.6 Comparing the number of negative collocates, negative FaC totals, and FaC totals as a % of total occurrences of CAUSE in the BoE and CKLC

	Number of Neg. Collocates		Neg FaC Total		Total Neg. FaC as a % of all occurrences		CKLC Difference
	BoE	CKLC	BoE	CKLC	BoE	CKLC	
N-4	18	3	1,112	58	1.2%	1.7%	+0.5
N-3	22	4	1,633	82	1.8%	2.4%	+0.6
N-2	15	7	1,864	120	2.1%	3.5%	+1.4
N-1	13	6	2,812	104	3.1%	3.0%	-0.1
N+1	26	12	8,688	260	9.7%	7.5%	-2.2
N+2	35	13	9,284	566	10.3%	16.4%	+6.1
N+3	26	7	4,245	315	4.7%	9.1%	+4.4
N+4	24	4	2,309	83	2.6%	2.4%	-0.2

Table 10.6 shows that the CKLC has lower FaC percentages at only three of the eight positions in the 4:4 span. Five of the eight are higher, and at least two are substantially higher. It appears, therefore, that the negative semantic prosody of CAUSE is at least as strong in the CKLC as it is in the BoE. A qualitative analysis of a random sample of 100 CKLC lines confirms this hypothesis, as all 100 use CAUSE to express negative propositions.

Most of the negative collocates of CAUSE in the CKLC are very general. These collocates (e.g. *accident*(*s*), *problem*(*s*), *disease*(*s*)) also appear in the BoE data, but in the BoE, as we have seen, there are many more specific collocates used. This difference could be because of the nature of the writing prompts given (general topics may beget general vocabulary, timed writing may not allow for dictionary searches of more specific words, etc.). It could also be simply because the Korean students' active vocabulary is not yet fully developed.

10.3 Phraseology in the CKLC and in the Classroom

This section discusses how the effects of phraseology on semantic prosody may influence teachers' and students' general approaches to meaning creation. It is difficult to come to any conclusions about phraseology *per se* in the CKLC because the frequencies of the phrases examined in this thesis are too low. As Table 10.7 below shows, only one of the phrases, *things happen*, occurs in the CKLC with a frequency high enough to facilitate detailed study.

The very low frequencies of these phrases in the CKLC recall Sinclair's (2003, p. 125) assessment of the relationship between phraseology and frequency:

> […] we may observe that a two-word phrase, while not as rare as the arithmetical predictions would have it, is still not nearly as common as the words that make it up, and a three-word phrase is even less common. […] So while happen to be is a phrase that is felt by native speakers to be quite normal and available, it is a lot less common

than happen to and a great deal less common than just happen (original underlining).

Table 10.7 Comparative frequencies in the BoE and CKLC of the phrases studied

	BoE		CKLC	
	freq	/mil	freq	/mil
things happen	1,497	3.34	24	4.54
these things happen	365	0.81	2	0.38
MAKE things happen	366	0.82	0	0
an accident waiting to happen	66	0.15	1	0.19
a distaster waiting to happen	60	0.13	0	0
the worst thing that can happen	64	0.14	0	0
the best thing that can happen	18	0.04	1	0.19

For example, the wordform *happen* occurs in the BoE 43,759 times, the two-word phrase *happen to* only 7,366 times and *happen to be* 1,233 times. In most corpus studies of phraseology the point is emphatically made that bigger is better when it comes to corpus size (see Hunston and Francis 2000; Sinclair 1991:18). Even the *ca.* 100-million-word British National Corpus was deemed an inappropriate general corpus for the investigations in this thesis because it is too small; the BNC contains, for example, only ten instances of *the ADJ thing that can happen*, and only seventeen occurrences of *a NOUN waiting to happen*. These are hardly enough to support to any strong conclusions. Similarly, the low frequencies of these and other phrases in the CKLC make it difficult to delve into how phraseology affects the semantic prosody in L2 English. However, results of investigations into BoE and enTenTen13 data do appear to have pedagogic significance.

10.3.1 Selecting a canonical form of *the* ADJ *thing that can happen*

This section discusses the pedagogical significance of teaching a canonical form of the semi-preconstructed phrase *the ADJ thing that can happen*. In the preface to the OSTI Report (Sinclair, Jones and Daley, 2004, p. xxiv), Sinclair frames at least one discussion of canonical

forms in terms of "language teaching" and the importance of such forms to students. In the context of *the ADJ thing that can happen*, the data presented in Chapter 8 suggests that if a student were presented with only lists of attested positive and negative adjective collocates in the absence of frequency data, he or she might reasonably infer that pragmatic choices to evaluate positively or negatively, or indeed to eschew evaluation altogether, are made equally frequently. As Table 10.8 shows, there are roughly equal numbers of negative, positive, and neutral adjectives attested in the phrase *the ADJ thing that can happen* in the enTenTen13. In fact, if we assume that *worth* and *sorst* are misspellings of *worst*, and if we conflate the misspelled *cruelest* with the correct *cruellest*, the numbers of negative and positive collocates attested in the EnTenTen13 are exactly the same.

Table 10.8 Fifty-three adjectives attested in the phrase *the ADJ thing that can happen* in the Ententen13, listed by evaluative polarity

	Negative			Positive			Neutral		
	Collocate	FaC	FaC % of Total	Collocate	FaC	FaC % of Total	Collocate	FaC	FaC % of Total
1	worst	1275	67.0%	Best	398	20.9%	only	44	2.3%
2	worse	70	3.7%	greatest	18	0.9%	other	6	0.3%
3	scariest	10	0.5%	easiest	4	0.2%	same	6	0.3%
4	bad	4	0.2%	nicest	3	0.2%	next	5	0.3%
5	saddest	2	0.1%	sweetest	3	0.2%	last	4	0.2%
6	hardest	2	0.1%	coolest	2	0.1%	first	2	0.1%
7	ugliest	2	0.1%	luckiest	2	0.1%	second	2	0.1%
8	riskiest	2	0.1%	Rarest	2	0.1%	highest	2	0.1%
9	stupidest	2	0.1%	optimum	1	0.1%	entire	1	0.1%
10	worth*	2	0.1%	noblest	1	0.1%	closest	1	0.1%
11	weirdest	2	0.1%	kindest	1	0.1%	kinkiest	1	0.1%
12	nastiest	1	0.1%	healthiest	1	0.1%	smallest	1	0.1%
13	darkest	1	0.1%	biggest	1	0.1%	possible	1	0.1%
14	severest	1	0.1%	strongest	1	0.1%	hottest	1	0.1%
15	cruelest	1	0.1%	finest	1	0.1%			
16	cruellest	1	0.1%	better	1	0.1%			
17	horriblest	1	0.1%	ideal	1	0.1%			
18	startling	1	0.1%	simplest	1	0.1%			
19	sorst*	1	0.1%						
20	extreme	1	0.1%						
21	strange	1	0.1%						
	Total	1383	72.7%	Total	442	23.2%	Total	77	4.0%

Sinclair (2004, p. xxiv) comments in the interview prefacing the OSTI report: "Initially, only the canonical form would be learned as a lexical item; the students would learn to recognize variants by themselves, and if uncertain about the meaning of a given variant, would have a dictionary to look it up." The canonical form is the variant that is selected most frequently, because, as Conrad (2004, p. 295) argues: "Something that occurs a thousand times is likely to be more use to a learner than something that just occurs a few times." This type of frequency data suggests that instead of lists of potential substitutions, it is likely more helpful for students to be given the information that, despite these similar numbers of collocates, the pragmatic decision to evaluate negatively via *worst* (1,275 occurrences) is made at least three times as frequently as the decision to evaluate positively via *best* (398 occurrences). As such, *the worst thing that can happen* could be called the canonical form as defined by Sinclair.

However, the potential importance to the learner of frequencies of more than one variant cannot be understated. For example, Sinclair (1991, p. 121) concludes his collocational study of *back* writing: "All the evidence points to an underlying rigidity of phraseology, despite a rich superficial variation." The same could be said of *the ADJ thing that can happen*. Despite the rich superficial variation of eighteen negative, eighteen positive, and fourteen neutral collocates attested in the enTenTen13, a single collocate stands out at the top of each list — *worst*, *best*, and *only* — which strongly suggests that the variants are neither superficial nor is the phrase itself entirely rigid. The rigidity is lessened somewhat by the observation that one lexical item stands out as, by far, the most frequent choice for each initial pragmatic evaluative decision.

Furthermore, the often subtle semantic differences among lower frequency variants could form the basis of a fruitful classroom activity or discussion. Sinclair (2004a, p. 281) writes:

> Most lexical items will include a substantial range of variation in their make-up, which keeps the management of variation in the hands of the teacher. In some circumstances the variation can be explored as a teaching point in itself, showing the nuances of meaning that can be created, the limits of the alternatives, and the possibilities of exploitation of the structure to create ironies and figures of speech.

It would no doubt be helpful for students to recognize, for example, that the collocates in each column of Table 10.8 are not necessarily synonyms, and therefore items in the lists of 'good' or 'bad' adjectives are not always easily interchangeable; the selection of a more specific adjective often adds a very precise element of meaning to the propositions. In the examples from the enTenTen13 below, *worst* is easily substitutable for *saddest* and *scariest* without sacrificing much of the intended meaning. However, students could be asked a) to discuss why *saddest* and *scariest* might be considered more specific choices than *worst* in these instances, and therefore why they are, perhaps, not interchangeable in these propositions; and b) to suggest alternative choices or other substitutable phrases.

2. This is clear when you realize that **the saddest thing that can happen** to a human being is the death of their child
3. A diagnosis that your child is sick or disabled is **the scariest thing that can happen** to a parent.

Furthermore, it is perhaps pedagogically significant that, taken together, the collocates *worst* and *best* combined account for 90.1% of occurrences of *the ADJ thing that can happen* in the BoE (70.3% and 19.8% respectively). Five adjectives comprise the remaining 9.9%. Similarly, in the enTenTen13 *worst* and *best* account for 87.9% of occurrences, and fifty-one adjectives comprise the remaining 12.1% of occurrences. This sharp drop in frequencies between the top two and third-most-frequent collocates appears to indicate that *best* is more than 'just' the second-most-frequent collocate. Put another way, when bad/unfavourable meaning is intended, *worst* is selected 92.4% of the time (1,275 times out of the total 1,380 negative occurrences). Similarly, when the pragmatic choice is to express a good/favourable proposition, *best* is

selected 90.0% of the time (398 times out of the total 442 positive occurrences). The dominating percentages of single collocates representing each pragmatic choice makes it possible to propose that there are in fact two canonical forms, one for each pragmatic choice.

This is not an argument that in every case the second-most-frequent collocate must be considered for co-canonical status. In this profile, though, the fact that *best* is both the only other adjective even close in frequency to *worst* and is of opposite evaluative polarity, indicates the potential significance, pedagogically, of presenting two co-canonical forms to students. That is, from a pedagogical perspective it might make sense to address two forms as, somewhat paradoxically, canonical.

10.3.2 Selecting a canonical form of the *a|an NOUN waiting to happen*

This section discusses the potentially problematic nature of selecting a canonical form of the semi-preconstructed phrase *a|an NOUN waiting to happen*. Sinclair notes that "[i]t happens that in most cases of varying realisations of a phrase, one of the alternatives is far more frequent than any of the others, and an obvious candidate for a canonical form, easy to teach and with the authority of a corpus behind it." (Sinclair, 2004a, p. 275). The collocates list in Table 8.8 below illustrate perhaps why Sinclair hedges this statement with 'in most cases' and does not seem to adopt an absolute approach to canonical form. The frequencies in Table 8.8 do not appear to support any but the most dogmatic of arguments for a frequency-based canonical form of the phrase. Although *accident* is the most frequently attested collocate in the BoE, it outnumbers *disaster* by only six occurrences (just over 3 percentage points). Similarly, *accident* in the EnTenTen13 outnumbers *disaster* by just twenty-five occurrences (0.5 percentage points).

The BoE and Ententen13 collocates and their relative frequencies in Table 8.8 are remarkably

similar. Seven of the ten collocates are identical and their frequencies vary by less than one percentage point in most cases. Most significant pedagogically is the fact that in both lists *accident* and *disaster* are the most frequent collocates, and their FaC values as percentages of the total number of occurrences are virtually identical. In the BoE they differ by only six occurrences, or 3.2 percentage points. In the enTenTen13 they differ by only twenty-five occurrences, equivalent to a difference of only 0.5 percentage points.

Table 10.9 A Comparison of the top-twenty most frequent attested noun collocates in the phrase *a|an NOUN waiting to happen* in the BoE and Ententen13

	Bank of English			Ententen13		
	NOUN	FaC	FaC % of Total Occurrences	NOUN	FaC	FaC % of Total Occurrences
1	accident	66	34.6%	accident	1,375	29.2%
2	disaster	60	31.4%	disaster	1,350	28.7%
3	tragedy	6	3.1%	lawsuit	140	3.0%
4	catastrophe	4	2.1%	tragedy	86	1.8%
5	incident	4	2.1%	catastrophe	83	1.8%
6	explosion	3	1.6%	nightmare	74	1.6%
7	injury	3	1.6%	problem	67	1.4%
8	breakdown	2	1.0%	injury	65	1.4%
9	champion	2	1.0%	explosion	35	0.7%
10	crisis	2	1.0%	adventure	34	0.7%
11	headline	2	1.0%	party	33	0.7%
12	lawsuit	2	1.0%	fire	30	0.6%
13	riot	2	1.0%	failure	28	0.6%
14	sacking	2	1.0%	headache	26	0.6%
15	winner	2	1.0%	crisis	24	0.5%
16	ambush	1	0.5%	scandal	22	0.5%
17	argument	1	0.5%	opportunity	21	0.4%
18	backlash	1	0.5%	mess	21	0.4%
19	bargain	1	0.5%	wreck	17	0.4%
20	collapse	1	0.5%	mistake	16	0.3%
	remaining 24 collocates	24	12.6%	remaining 558 collocates	1,392	29.6%
	Total	191	100.0%	Total	4,701	100.0%

It is also perhaps worth noting, however, that in both lists *accident* is more frequent. For additional comparison, the *ca.* one-hundred-million-word BNC was also queried. There, the phrase is found only seventeen times, and *accident* and *disaster* again dominate with nine and six occurrences respectively.

Evidence from The CCED (2001, p. 8) indicates that *accident* is the *de facto* canonical form, because *an accident waiting to happen* has its own entry: "If you describe something or someone as **an accident waiting to happen**, you mean that they are likely to be a cause of danger in the future, for example because they are in poor condition or behave in an unpredictable way" (original emphasis). There is no similar entry for *a disaster waiting to happen* despite its very high frequency in the BoE, which is the corpus used to inform Collins' definitions. Dictionary evidence alone would seem to suggest that *a disaster waiting to happen* is little more than a variation of *an accident waiting to happen*. Indeed, the CCED (2001, p. 431) defines *disaster* as "a very bad accident."

The similarity in frequency is significant pedagogically because it suggests that we can expect a student to encounter *a disaster waiting to happen* effectively as frequently as *an accident waiting to happen*. As such, teaching a single canonical form seems counter-productive. Recall that in the case of *the ADJ thing that can happen* it was suggested that the initial pragmatic choice to evaluate positively or negatively might be taken into account in a discussion — especially one that is pedagogically motivated — of a phrase's canonical form. It was argued that although the negative form is clearly chosen far more frequently, the positive (and perhaps even the neutral) form ought not to be ignored and is perhaps worthy of co-canonical status. In the case of *a|an NOUN waiting to happen*, co-canonical status is not based on the pragmatic choice of how to evaluate — evaluation is overwhelmingly negative — but rather on the very similar frequency data.

However, a further hypothesis is that *an accident waiting to happen* and *a disaster waiting to happen* are not, in fact, two variations of the same form. In Sinclair's (2004a, p. 271) discussion of "four aspects of the way we perceive and handle language that can be a nuisance in teaching

and learning," he writes:

> [B]ecause it is lexical, the item is sensitive to the overall meaning that is being created, so that if a choice that could be made at one place in structure would alter the meaning of the whole, it must indicate the presence of another lexical item which shares some of the same elements (Sinclair, 2004a, p. 283).

The question is, then, does the substitution of *disaster* for *accident* in this phrase substantively change the meaning of the phrase, thereby creating a discrete lexical item? Given the dictionary definition of *disaster* above, it would seem that the difference is a matter of degree, not semantic disparity, and comparison of collocational profiles of the two variants appears to support this assessment. Where the phrase acts as node, as Table 10.10 illustrates, collocates in the 4:4 span by frequency for *a|an accident|disaster waiting to happen* are very similar. At N+1 forty of the top fifty most frequent collocates in each profile are shared. The average is just over thirty-two shared collocates per position in the PFT.

Table 10.10 Illustration of whole phrases acting as node

N-4	N-3	N-2	N-1	NODE	N+1	N+2	N+3	N+4
stuff	is	nasty	and	an accident waiting to happen.	Get	an	old	refridgerator
had	not	been	prevented.	a disaster waiting to happen,	in	a	country	which

However, these shared collocates are primarily functional/grammatical. Further research into lexical collocates may reveal important differences or similarities. If these two phrases were separate, though nearly synonymous, we might expect them to be primed quite differently: "Co-hyponyms and synonyms differ with respect to their collocations, semantic associations and colligations" (Hoey, 2005, p. 13). Their similar collocational profiles suggest that they are not two nearly-synonymous phrases, but rather a single phrase structure demonstrating only slight internal variation.

One final piece of evidence suggesting that *a|an accident|disaster waiting to happen* are

variants of the same semi-preconstructed phrase and not discrete lexical items is that they both appear to be "primed to occur in […] certain positions within the discourse; these are its textual colligations" (Hoey, 2005, p. 13) . In the enTenTen13, *an accident waiting to happen* is found in clause terminal position in *ca.* 77% of instances; *a disaster waiting to happen* ends a clause in *ca.* 82% of instances.

10.4 Grammatical Patterning in the Classroom

This section looks briefly at the pedagogical implications of the effects of grammatical patterning on the semantic prosody of CAUSE. As discussed above, it is difficult to compare patterning in the CKLC because the tokens are not tagged for parts of speech. However, isolating *to* at N+2 of the occurrences of CAUSE in the CKLC allowed for the creation of a small concordance of the pattern ***CAUSE*** **n to-inf**. As before, this is a basic version of the pattern evincing only the simplest noun groups, but it serves to indicate how awareness of semantic prosody in general and grammatical patterning associated with an item can be beneficial to students.

In all, there are forty-four unique verbs at N+3; Table 10.11 shows the top ten, their FaC values, and the FaC percentage of the total occurrences.

Table 10.11 Top-ten verbs by frequency at N+3 for seventy-six lines of the pattern *CAUSE* n to-inf in the CKLC

Collocate	FaC	%
relax	12	15.8%
get	6	7.9%
feel	5	6.6%
make	4	5.3%
have	4	5.3%
fail	2	2.6%
be	2	2.6%
think	2	2.6%
smoke	2	2.6%
suicide[6]	2	2.6%
Total	41	53.9%

These collocates appear to show a high degree of smoothing; only two are negative, one is positive, and seven are neutral or unknown. However, many of the neutral verbs appear to require a completion of one kind or another: there are examples verbs that are potentially delexical (*get*, *make*, *have*), linking (*feel*, *be*), or reporting (*think*).For example, the verbs *get*, *feel*, and *make* in the following lines are all part of phrases that evaluate negatively.

4. the physical punishment could **cause** children to get hurt
5. If you swear to someone, it **causes** someone to feel bad but also
6. words that media says. This might **cause** voters to make wrong decisions.

In fact, qualitative examination of the concordance reveals a very strong negative prosody. Table 10.12 shows that fifty-eight of the seventy-six instances of the pattern (76.3%) contribute to negative propositions.

Table 10.12 Results of qualitative examination of seventy-six instances of *CAUSE* n to-inf in the CKLC

	Freq	%
Negative	58	76.3%
Positive	15	19.7%
Neutral/Unknown	3	3.9%
Total	76	100.0%

The positive instances are something of an anomaly. As Table 10.11 above has shown, *relax* is the most frequent collocate at N+2 and accounts for twelve of the fifteen positive instances of the pattern. These instances of *relax* all relate in the corpus to the effect of colour on mood, and it is clear that a very specific writing prompt has instigated all of them. This collocative clash could inspire a classroom discussion or lesson on appropriate usage of the semantic prosody of CAUSE in general and how this pattern tends to smooth (but not reverse) this prosody.

As a final brief note, even though ***CAUSE by*** **n** perhaps does not smooth the negative semantic prosody of CAUSE enough to warrant special consideration in the classroom, it remains a pattern

worth noting. As discussed above in Section 10.2.1, *by* is not the most frequent collocate of CAUSE in the CKLC. This contrasts with the BoE data may be worth exploring in detail. It is likely not helpful to adopt a prescriptive stance regarding this pattern, but a comparison of frequent noun phrases CAUSE is often found in (e.g. *the cause of*) and the frequent verb-complement phrases discussed here, including by not limited to ***CAUSE by*** **n** may be helpful to the student.

10.5 Conclusion

It is argued in Section 10.2 above that despite the lower numbers of negative collocates in the CKLC their comparable FaCs indicate that the Korean students have a strong grasp of the negative prosody of CAUSE. For example, at N+1 of the BoE there are more than double the number of negative collocates than at N+1 of the CKLC (26:12); at N+2 the ratio is even more distinct (35:13). However, the FaC totals are comparable: the negative collocates at N+1 and N+2 of the BoE Picture account for 9.7% and 10.3% of the 89,830 instances of CAUSE respectively, compared to 7.5% and 16.4% in the CKLC. This indicates that the pragmatic meaning of CAUSE is 'known' to the students, though likely unconsciously, but compared to native speakers, the vocabulary at their command to express this pragmatic meaning is limited.

Section 10.3 discusses the potentially problematic process of selecting a canonical form of the semi-preconstructed phrases analysed in this thesis. The phrase *the ADJ thing that can happen* has a potentially pedagogically useful collocational profile. It is argued that, even where one paradigmatic selection rises to canonical status by virtue of frequency, e.g. *worst*, the evaluative polarity of the most frequent collocates play a pedagogically useful role. Specifically, it is suggested that when paradigmatic choices of opposite evaluative polarity are both relatively

frequent and salient, then co-canonical status might be a helpful notion.

In a related discussion, it has been shown that the phrase *a|an NOUN waiting to happen* also has two dominant collocates. In this case polarity is not an issue, *i.e.* the pragmatic/functional choices are virtually always negative. The difficulty in this case lies in the virtually identical frequencies of *accident* and *disaster* in the data, and the interesting fact that despite the very close FaC values, *accident* appears to always be the most frequent. Where paradigmatic choices are possible, Sinclair argues "the paradigms are prioritised on frequency grounds" (Sinclair, 2004a, p. 283). However, when the paradigms are this close in frequency, it would seem prudent for both (or all) to be given to students.

Finally, Section 10.4 has briefly discussed the value of investigating grammatical patterning in the language classroom. Specifically, though the patterns are neither as clear (because of limits in student grammatical accuracy) in the CKLC nor as available for inspection (because the corpus is not tagged for parts of speech), even a very small concordance can reveal trends that students might benefit from observing.

Notes

[1] Details of the Combine Korean Learner Corpora used in the current study:

- I compiled the **Chungbuk National University (CBNU) Corpus of Written English** between September 2014 and March 2016. Texts are paragraph-length, general English compositions written by Korean undergraduate university students.
- The **Chongshin University Corpus of Written English** comprises Korean undergraduate university students' writing collected by Heidi Nam and submitted to me, with permission of the students, for the purposes of this study.

- The **Gachon Learner Corpus** (Carlstrom and Price, 2012), is a collection of 2,500 short responses (100-150 words each) to twenty questions. The corpus is available online at http://koreanlearnercorpusblog.blogspot.kr/p/corpus.html

- The **International Corpus Network of Asian Learners of English** (**ICNALE**) comprises short essays written by university students. Only the Korean sub-corpus has been used employed in the current study. The corpus is available for download at http://language.sakura.ne.jp/icnale/icnale_online.html.

- The **Neungule Interlanguage Corpus of Korean Learners of English** (**NICKLE**) comprises a variety of university student essays (descriptive, narrative, argumentative, etc.) written in English. Access to the corpus can be granted by contacting the its creator and administrator, Ji-Myoung Choi at amancio.choi@gmail.com.

- The **Yonsei English Learner Corpus** (**YELC**) comprises short (100-300 words) written descriptive and argumentative texts written in English by undergraduate Korean students. It is available for research purposes by contacting CK Jung at corpuslab@yonsei.ac.kr.

[2] Antconc is corpus analysis freeware available for download at http://www.laurenceanthony.net/software/antconc/

[3] Calculated by Antconc.

[4] This is the "Final Version" downloaded from the creators' website: http://koreanlearnercorpusblog.blogspot.kr/p/corpus.html. Sometime after this version was made available for download, The Gachon Korean EFL Learner Corpus, a ca. 1.8 million-word web version became available at: https://corpling.uis.georgetown.edu/cqp/gachon/index.php?thisQ=who_the_hell&uT=y

[5] In addition to part-of-speech differences, the overuse of CAUSE in the learner corpus may in part be due to the nature of one or more of the writing prompts the students were given. One of the CBNU prompts, for instance, was "[w]rite a short cause-effect essay on the topic 'obesity'." This, of course, also begins to explain why *obesity* is found in all but N+4 of the 4:4 span of cause in the CKLC, and is ranked twenty-fourth in the T-list for the CKLC.

[6] In these instances, *suicide* is being used as a verb, which is not uncommon in Korean students writing in English.

CHAPTER 11:
SUMMARY OF RESULTS AND SUGGESTIONS FOR FURTHER RESEARCH

11.1 Introduction

In this thesis, I have examined the semantic prosodies of two high-frequency verbs, CAUSE and HAPPEN. Corpus data contains evidence that these prosodies are significantly affected by at least four linguistic factors, which are summarized in this chapter. First, Section 11.2.1 summarizes the effects of collocation on the semantic prosodies of the lemmas CAUSE and HAPPEN; Sections 11.2.2 and 11.2.3 review the effects of phraseology on the semantic prosody of HAPPEN; and 11.2.4 provides a summary of the smoothing effects of specific grammatical patterning and register on the semantic prosody of CAUSE. Finally, Section 11.3 discusses some of the limitations of the current study and offers suggestions for further research.

11.2 Summary of Results

11.2.1 Collocation

The current study shows that semantic prosody as a collocational phenomenon is problematic for a number of related reasons. First, as McEnery and Hardie (2012, p. 123) argue, "as soon as we […] attempt to pin down collocation either operationally or conceptually, we find a great multitude of different definitions." The literature review in Chapter 2 focused on Partington's (1998, pp. 15–16) three classifications of collocation—textual, psychological, and statistical. Although this is a corpus-based study, and would therefore seem to benefit from a statistical/frequency-based definition of collocation, it was shown in Chapter 6 that such an approach is itself problematic in the context of studies of semantic prosody.

Specifically, it has been shown that statistical significance does not appear to be a requirement

of individual collocates evincing an item's semantic prosody. First of all, different statistical measures return sets of ostensibly "significant" collocates that are often very different. For example, the top five collocates in the BoE T-List for CAUSE are *by*, *problems*, *damage*, *which*, and *can*, while the top five collocates in the BoE MI-List for CAUSE are *sulphurous*, *miscompute*, *diazoxide*, *cryptosporidiosis* and *ducreyi*. Reasons for these differences have been discussed at length above and elsewhere (*cf.* McEnery, Xiao, and Tono 2006:215–20). More importantly, perhaps, is the observation that the frequency of an individual evaluative collocate is not a significant factor in observations of semantic prosody. This is primarily because semantic prosody is activated not by a single collocate but by groups of collocates that evaluate similarly. Therefore, a more appropriate metric of significance is the total frequency (FaC) of all evaluative collocates observed in a set. For example, in Section 7.1.1, the collocational profile of the phrase *things happen* was observed to contain an almost equal number of negative and positive collocates (a ratio of 14:15), but the total FaCs reveal that the negative collocates are selected almost 30% more frequently (179:128), suggesting a somewhat stronger negative prosody.

Furthermore, qualitative analyses throughout this thesis also repeatedly illustrate the argument that collocates — single-word or structurally complex (i.e. phrasal) — must exhibit a clear syntactic and semantic relationship to the node in order to be considered exemplars of that node's semantic prosody. The primary significance of this claim is that in instances where single-word or phrasal evaluative collocates are observed to have little or no direct syntactic or semantic connection to the node, evaluation is said to be "textual" rather than prosodic (*cf.* Mahlberg 2005). It has been argued that this distinction between modes of evaluation and the evidence that counts for each mode may be the source of the misperception that HAPPEN has a

strong negative semantic prosody. Certainly, the environment of HAPPEN does contain a great deal of evidence of evaluation, but this evidence is not likely to appear on collocational profiles that display only single words found in the 4:4 span.

This leads to the final, unforeseen, conclusion of the collocational analysis of HAPPEN, namely that it does not appear to have a negative semantic prosody at all, despite frequent claims to the contrary (Sinclair, 1991, 2003; Bublitz, 1996; Partington, 2004, 2014). It is only when specialized corpus queries are employed (to isolate noun collocates in the 4:0 span, for example) that any clear evidence of negative semantic prosody is evident. These focussing queries, however, return only a very small fraction of the total occurrences of HAPPEN and, as such, appear to evince only the most superficial evidence of semantic prosody. Although 27 of the top 50 noun collocates returned at N-1 are negative (see Table 6.9), the Total FaC values for these negative collocates account for only 2,076 instances, or just 1.4% of the total occurrences of HAPPEN in the BoE. Similar queries for noun collocates at N-2, N-3, and N-4 return even fewer negative collocates.

11.2.2 Phraseology part one

Investigation into the effects of phraseology on semantic prosody began with the "nested" collocation (*cf.* Hoey 2005) *things happen*. Comparative analysis of the numbers of positive and negative collocates and their combined FaCs has been summarized in the previous section (11.2.1). Additional analysis has revealed, however, that the phrases *good things happen* and *bad things happen* can have their transparent/'core' evaluative polarities reversed with the addition of *to bad people*. That is, the phrase *good things happen* evaluates positively, but *good things happen to bad people* is negative. Likewise, *bad things happen* is negative¸ but *bad*

things happen to bad people evaluates positively.

Investigation into *these things happen* revealed three distinct meanings of the phrase. This phrase used in a literal, open-choice, construction was counted only twelve times in the BoE data, seven of which were observed to evaluate negatively. However, only four of these twelve showed evidence of semantic prosody. A second meaning, "this type of thing happens", was noted sixty times in the data. Of these, thirty were observed to evaluate negatively, but only six of these thirty evinced semantic prosody; the remainder evaluated textually. Finally, analysis of a random 100-line sample of the 288 lines expressing ameliorative meaning of the phrase showed that eighty-three are negative, but only twenty-seven of these evaluated via semantic prosody, and the remainder were textual.

Only one sense of the phrase MAKE *things happen* was found in the BoE data. A person or group who *makes things happen* creates or finds opportunities to do something beneficial, usually in a very specific arena such as sport or business. Notably, this phrase was found to evaluate positively, which is somewhat unexpected given the ostensible negative semantic prosody of HAPPEN; only five of the 100-line sample concordance were found to evaluate negatively. Of the ninety-five remaining lines, only six were shown to evaluate via positive semantic prosody (i.e. where the thing that is made to happen is observed in a direct semantic and syntactic connection to the node), eighty-three evaluated via positive textual meaning, and six remained unknown.

11.2.3 Phraseology part two

Examination of the effects of phraseology on semantic prosody continued in Chapter 8 with analyses of the phrases *a*|*an* NOUN *waiting to happen* and *the* ADJECTIVE *thing that can*

happen. The chapter began by comparing collocational profiles of the phrases as they increased in size. Comparison of T-pictures for iterations of these phrases as they increase in size showed in both cases increasing evidence of negative semantic prosody. That is, the profile for *waiting to happen* contains more occurrences of negative collocates than the profile for *to happen*, which, in turn, contains stronger evidence of negative semantic prosody than the lone word form *happen*. Similarly, the T-picture for *thing that can happen* shows higher negative FaC totals than *that can happen*, which is also show higher values than collocates of *can happen*.

The primary significance of this observation is that, in both cases, it appears that characterizing the most frequently occurring evaluative words in the profile as "collocates" is potentially misleading. Corpus data shows that these evaluative words — *worst* and *best, accident* and *disaster* — are frequent in the profiles almost entirely because they occur in these longer phrases and that co-selection in any other phrase or open-choice variant is exceedingly infrequent.

It was also noted that the phrases themselves do not have negative semantic prosodies in the sense the they do not have considerable numbers of negative collocates in the 4:4 span, and the negative collocates that are found in the profile are selected very infrequently. This is not to suggest that these phrases do not evaluate, however. It has been argued that these two phrases both evaluate via their core meaning depending on the adjective or noun selected.

The final phraseological effect observed in Chapter 8 was that phrases that appear to be very similar can, in fact, evaluate quite differently. It was shown that *the worst **thing** that can happen* is used in a literal sense more frequently and in ameliorative sense much less frequently than *the worst that can happen*. It was argued that this is an effect of the ostensibly neutral general

noun *thing* acting as evaluation carrier (*cf*. Hunston and Francis 2000; Mahlberg 2005).

11.2.4 Grammatical patterning and register

The investigation into the effects of semantic prosody and register was directly inspired by a short concordance of neutral instances of CAUSE taken from the **newsci** (New Scientist) sub-corpus of the BoE used by Hunston (2007) to illustrate that not only that the semantic prosody of CAUSE is not absolute, but also that it appears to be activated primarily when related to human beings and their endeavours. As noted above, however, this twelve-line concordance comprises only two grammar patterns of CAUSE. Chapter 9, therefore, examined whether the smoothing effects observed in the concordance were an effect of these patterns, the **newsci** register, or both.

First, investigation of the lemma CAUSE in the full BoE compared to five specific registers showed convincing evidence that the negative prosody is smoothed extensively in **usacad**, somewhat less in **books**, but very little in **news** (in fact, the negative semantic prosody of CAUSE is slightly stronger in **news** than it is in the BoE), **spoken**, or **newsci**. This result is somewhat surprising given that the lines selected for Hunston's illustration all came from **newsci**. It was argued in Chapter 9 that this lack of smoothing may be because of the fact that **newsci** effectively straddles at least two registers, namely news reporting and academic/scientific writing. It was argued that in Hallidayan terms, although the field and mode of **usacad** and **newsci** might be considered quite similar (i.e. scientific/academic writing), in fact **newsci** aims to communicate to broader audiences — more akin to the **news** register — meaning that the tenors of **newsci** and **usacad** are significantly different.

Four grammatical patterns of CAUSE were then investigated in the full BoE and **newsci** to determine whether these patterns had an observable effect on its semantic prosody. Results

showed no evidence of smoothing for the patterns **CAUSE n** or **CAUSE n n** in either investigation. Some evidence of smoothing was observed for the pattern **CAUSE *by* n** the BoE, and this smoothing appears slightly stronger in **newsci**; qualitative analysis showed that a difference of *ca.* 15 percentage points (BoE 87.4% negative, newsci 72.0% negative). There is strong evidence of prosodic smoothing for the pattern **CAUSE n to-inf** in both the BoE and **newsci**, and again, the smoothing is more prevalent in **newsci** by about 10 percentage points (52.5% negative in the BoE, 42.2% negative in **newsci**).

Interestingly, it was observed that many of the nouns selected in the pattern **CAUSE n to-inf** appear to refer explicitly to human beings, which is somewhat surprising because Hunston (2007) hypothesizes that CAUSE has a negative semantic prosody only when it refers to humans and human concerns. She suggests that the prosody is deactivated when that which is caused is inanimate and is not related to human endeavour. That this pattern both smooths and often concerns humans requires further investigation.

Finally, as Hunston (2007, p. 263) suggests, "we might argue that particular registers select one lexical phenomenon more frequently than another." It was indeed demonstrated that the two patterns shown to have a smoothing effect on the semantic prosody of CAUSE are selected more frequently in **newsci** and **usacad**, although further research is required to determine the distribution of the relevant CAUSE patterns in the various registers.

11.2.5 Pedagogical implications

Chapter 10 discussed some pedagogical implications and applications of the collocational, phraseological, and grammatical patterning observations made throughout this thesis. First, it was noted that, somewhat surprisingly, Korean student writers appear to activate the negative

semantic prosody of CAUSE about as often as L1 writers. Collocational data from a learner corpus of Korean university student writing was compared to data from the BoE. It was observed that although there were notably fewer negative collocates of CAUSE in the students' writing, the negative FaC total (the total number of times the negative collocates were selected) was comparable to the BoE negative FaC total. It was argued that the pragmatic/evaluative meaning activated by the students despite the more limited vocabulary choices to express that meaning.

The studies of the phraseological behaviour of HAPPEN led to a discussion of the importance of teaching canonical forms of phrases. It was argued that the fact that *worst* and *best* are the first and second most frequently selected adjective collocates for the phrase *the* ADJECTIVE *thing that can happen* suggests that it may be pedagogically useful to grant co-canonical status to these variants. The point of such a discussion in the classroom would be to emphasize that although the most frequent selection is negative, a positive choice is also very frequent and available to the user. A simple frequency-based canonical variant would neglect this pedagogically useful information.

The phraseological behaviour of *a|an* NOUN *waiting to happen* also lends itself to a useful classroom discussion. In this case, there are two very frequent negative collocates — *accident* and *disaster* — the frequencies of which are so close, co-canonical status would, again, seem prudent pedagogically. Here, the classroom discussion would change its focus from polarity to frequency, specifically the fact that even though the phrases appear to be virtually interchangeable and are remarkably close in frequency in the corpora checked, *accident* is consistently the most frequently selected collocate, but only by a very small margin.

Finally, Chapter 10 briefly discussed the potential utility of grammatical patterning discussions in the classroom. It is difficult to generalize from the student data because a) student grammar is often "incorrect" and therefore interrogations of learner corpora for specific patterns are likely to return smaller concordances, and b) the learner corpus used in the current study is not tagged for parts of speech, making searches especially challenging (this could be fairly easily remedies in further studies, however). It was argued that despite these limitations, even small concordances of student writing can be useful in the classroom; samples could be tagged/labelled/arranged manually in ways that could benefit students. It was shown, for example, that *by* is not the most frequent collocate of CAUSE in the CKLC as it is in the BoE. This suggests a fundamental difference in the ways CAUSE is used by students, and the reasons for this difference could form the basis of a fruitful classroom discussion.

11.3 Limitations of the Current Study and Suggestions for Further Research

The current study is limited in at least three ways, and so observations reported here should be considered largely preliminary. First, the study has examined a very limited number of items: only two lemmas (CAUSE and HAPPEN), five phrases associated with HAPPEN, and four grammatical patterns of CAUSE have been investigated. Although corpus data do clearly show how, in each case, semantic prosody is affected by collocation, phraseological behaviour, register, and grammatical patterning, this remains a very small sample.

Future study would no doubt benefit from study of more lemmas and individual word forms where appropriate. For example, PROVIDE has been shown to have a strong positive prosody (Stubbs, 2001b, p. 24). However, during a pilot study in preparation for this thesis, I noticed an apparent collocational clash in the ubiquitous F.B.I. anti-piracy warning shown at the beginning

of films on VHS tapes, DVDs, and Blu-ray discs. The warning reads: "Federal law **provides severe civil and criminal penalties** for the unauthorized reproduction, distribution, or exhibition of copyrighted motion pictures, video tapes or video discs" (emphasis added). It would be interesting to investigate whether it is the legal register, or some other phraseological or grammatical factor reversing the expected positive polarity of PROVIDE in this example.

Additionally, the phraseological behaviour of HAPPEN is a potentially rich area for further study. Salient MWUs and other phrases associated with HAPPEN are plentiful, and corpus investigation of more phrases would certainly complement the current study and perhaps add new insights into how these phrases affect the semantic prosody of HAPPEN.

Secondly, only a small number of registers were investigated in the current study. Examinations focussed on the **newsci** sub-corpus of the BoE because the lines selected by Hunston in her seminal discussion of the problematic nature of semantic prosody came from the same sub-corpus. However, results of the current study showed stronger prosodic smoothing effects in the more purely academic register **usacad**. Further study would benefit from looking into more specific scientific and academic registers and genres. For example, there may be observable differences in how CAUSE and its grammar patterns (not to mention other lemmas and associated phrases and patterns) evaluate specifically in the biological, physical, and chemical sciences, for example. Indeed, any number of registers may be observed to have a smoothing effect on CAUSE and other items.

Thirdly, the Korean learner corpus compiled for the current study is untagged for parts of speech. This limitation makes querying the corpus and directly comparing data very time consuming and creates a risk that salient data will be overlooked. Every effort has been made to analyse

and present learner data presented here as closely and accurately as possible, but further research employing a tagged corpus would no doubt provide more detailed accounts. Additionally, learner corpora from other languages would complement the current results and no doubt reveal as yet unobserved effects.

11.4 Conclusion

Semantic prosody instantiates the idiom principle itself (Hunston, 2002, p. 142). The notion that language is essentially phraseological — "People do not speak in words; they speak in phrasemes" (Mel'čuk, 1995, p. 169) — is one that has quickly come to be accepted as axiomatic among Applied Linguists, but has yet to fully take hold in the realms of language education. Materials designers, teachers, and students have now begun to embrace the phraseological essence of meaning creation, but a great deal more needs to be done to bring a phraseological model of language to students. This makes further investigations of semantic prosody and related realms of meaning creation of crucial importance as we move forward.

Corpus data has shown that semantic prosody is a complex phenomenon that involves much more than merely counting positive and negative single-word collocates in the immediate vicinity of the node word. On the contrary, evidence presented in this thesis has shown that collocation, phraseological behaviour, grammatical patterning, and register can significantly influence observations of semantic prosody. Semantic prosody has been shown to be activated by structurally complex (phrasal) collocates, at times found a far from the node. It has also been demonstrated that semantic prosody is perhaps best conceived on a cline of evaluative meaning between explicit, core, evaluative meaning and textual meaning that relies on extra linguistic or contextual knowledge to be activated. It has been shown that certain grammatical patterning

can have a significant effect how an item evaluates, and that certain registers do appear to smooth certain prosodies.

Perhaps more than this, though, this study of semantic prosody has led to some unexpected observations. First is the suggestion that collocation itself, at least in some instances, is illusory and that words that may appear to be significant co-selections in fact occur in profiles solely by virtue of the phraseological behaviour of the node and not because of other factors that have been shown to affect collocation (*cf.* Walker 2011a, 2011b). Furthermore, results of phraseological behaviour have also suggested that the notion of a canonical variation of a phrase (Sinclair, Jones and Daley, 2004) is also potentially illusory. For some highly variant phrases, only a quasi-canonical iteration appears. Corpus data shows that selecting a canonical form is not always a simple matter of having a computer pluck the most frequent variant from masses of computer data. In contrast, canonical status is often 'fuzzy' and requires human intervention, interpretation, and explanation.

This study originally intended to show only that the semantic prosodies of two high-frequency verbs, CAUSE and HAPPEN, were either smoothed, reversed, or strengthened by four linguistic factors. In every case, evidence has shown these effects. However, these results and the 'extra' unexpected observations, require a great deal more research in order to expand and develop them.

References

Adolphs, S. (2006) *Introducing Electronic Text Analysis: A Practical Guide for Language and Literary Studies*. New York: Routledge.

Altenberg, B. (1998) 'On the Phraseology of Spoken English: The Evidence of Recurrent Word-Combinations', in Cowie, A. P. (ed.) *Phraseology: Theory, Analysis, and Applications*. Oxford: Clarendon Press.

Angermuller, J., Maingueneau, D. and Wodak, R. (eds) (2014) 'The Discourse Studies Reader edited by', in *The Discourse Study Reader: Main currents in theory and analysis*. Amsterdam and Philadelphia: John Benjamins Publishing Company.

Barnbrook, G. (1996) *Language and Computers: A Practical Introduction to the Computer Analysis of Language*. Edinburgh: Edinburgh University Press.

Barnbrook, G., Mason, O. and Krishnamurthy, R. (2013) *Collocation: Applications and Implications*. Palgrave MacMillan.

Bednarek, M. (2008) 'Semantic preference and semantic prosody re-examined', *Corpus Linguistics and Linguistic Theory*, 4(2), pp. 119–139.

Bhatia, V. K. (1997) 'Genre-mixing in academic introductions', *English for Specific Purposes*, 16(3), pp. 181–195.

Biber, D. (1988) *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, D. (2009) 'A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing', *International Journal of Corpus Linguistics*, 14(3), pp. 275–311.

Biber, D. and Conrad, S. (2009) *Register, Genre, and Style*. Cambridge/New York: Cambridge University Press.

Biber, D., Conrad, S. and Cortes, V. (2004) 'If you look at ...: Lexical Bundles in University Teaching and Textbooks', *Applied Linguistics*, 25(3), pp. 371–405.

Biber, D., Conrad, S. and Reppen, R. (1994) 'Corpus-based Approaches to Issues in Applied Linguistics', *Applied Linguistics*, 15(2), pp. 169–189.

Bublitz, W. (1996) 'Semantic prosody and cohesive company: somewhat predictable', *Leuvense Bijdragen*, 85(1–2), pp. 1–32.

Burger, H. (1998) *Phraseologie: Eine Einführung am Beispiel des Deutschen*. Berlin: Erich Schmidt.

Carlstrom, B. and Price, N. (2012) *The Gachon Learner Corpus*.

Clear, J. (1993) 'From Firth Principles - Computational tools for the study of collocation', in

Baker, M., Francis, G., and Tognini-Bonelli, E. (eds) *Text and Technology: In Honour of John Sinclair*. Amsterdam and Philadelphia: John Benjamins Publishing Company, pp. 271–292.

Collins Cobuild English Dictionary for Advanced Learners (2001) *English Dictionary for Advanced Learners*. Third. Edited by J. Sinclair. Glasgow: Harper Collins Publishers.

Conrad, S. (2004) 'Corpus linguistics, language variation, and language teaching', in Sinclair, J. (ed.) *How to use Corpora in Language Teaching*. Amsterdam/Philadelphia: John Benjamins Publishing Company, pp. 67–85.

Cotterill, J. (2001) 'Domestic Discord, Rocky Relationships: Semantic Prosodies in Representations of Marital Violence in the O.J. Simpson Trial', *Discourse & Society*, 12(3), pp. 291–312.

Cowie, A. P. (1994) *Phraseology, The Encyclopedia of Language and Linguistics*. Edited by R. E. Asher. Oxford: Oxford University Press.

Cowie, A. P. (1998) 'Introduction: Past Achievements And Current Trends', in Cowie, A. P. (ed.) *Phraseology: Theory, Analysis, and Applications*. Oxford: Clarendon Press.

Danielsson, P. (2007) 'What Constitutes a Unit of Analysis in Language?', *Linguistik Online*, 31.

Dilts, P. and Newman, J. (2006) 'A note on quantifying "good" and "bad" prosodies', *Corpus Linguistics and Linguistic Theory*, 2(2), pp. 233–242.

Dudley-Evans, T. (1994) 'Genre analysis: An approach to text analysis for ESP', in Coulthard, M. (ed.) *Advances in Written Text Analysis*. London and New York: Routledge.

Ellis, N. C., Frey, E. and Jalkanen, I. (2009) 'The psycholinguistic reality of collocation and semantic prosody (1): Lexical access', in Römer, U. and Schulze, R. (eds) *Exploring the Lexis-Grammar Interface*. Amsterdam/Philadelphia: John Benjamins Publishing Company, pp. 89–114.

Erman, B. and Warren, B. (2000) 'The idiom principle and the open choice principle', *Text*, 20(1), pp. 29–62.

Flowerdew, J. and Dudley-Evans, T. (2002) 'Genre Analysis of Editorial Letters to International Journal Contributors', *Applied Linguistics*, 23(4), p. 463–489+550.

Goldberg, A. E. (2003) 'Constructions: A new theoretical approach to language', *Trends in Cognitive Sciences*, 7(5), pp. 219–224. doi: 10.1016/S1364-6613(03)00080-9.

Granger, S. and Paquot, M. (2008) 'Disentangling the phraseological web', in Granger, S. and Meunier, F. (eds) *Phraseology: An interdisciplinary perspective*. Amsterdam/Philadelphia: John Benjamins Publishing Company, pp. 27–49.

Gries, S. T. (2008) 'Phraseology and linguistic Theory: A brief survey', in Granger, S. and Meunier, F. (eds) *Phraseology: An interdisciplinary perspective*. Amsterdam/Philadelphia: John Benjamins Publishing Company, pp. 3–25.

Groom, N. (2009) *Phraseology and Epistemology in Academic Book Reviews: A Corpus-Driven Analysis of Two Humanities Disciplines, Academic Evaluation: Review Genres in University Settings*. Edited by K. Hyland and G. Diani. Palgrave MacMillan.

Halliday, M. A. K. (2014) 'Language as Social Semiotic', in *The Discourse Study Reader: Main currents in theory and analysis*. London: Arnold, pp. 263–271.

Halliday, M. A. K. and Hasan, R. (1976) *Cohesion in English*. London: Longman.

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

Hoey, M. (1997) 'From concordance to text structure: new uses for computer corpora', in Lewandowska-Tomaszczyk, B. and Melia, P. J. (eds) *PALC '97: Proceedings of Practical Applications in Language Corpora Conference*. Lodz: University of Lodz, pp. 2–23.

Hoey, M. (2005) *Lexical Priming: A New Theory of Words and Language*. London and New York: Routledge.

Hoey, M. (2007) 'Lexical priming and literary creativity', in Hoey, M. et al. (eds) *Text, Discourse and Corpora: Theory and Analysis*. London: Continuum.

Hoey, M. *et al.* (2007) *Text, Discourse and Corpora: Theory and Analysis*.

Hunston, S. (2002) *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Hunston, S. (2007) 'Semantic prosody revisited', *International Journal of Corpus Linguistics*, 12(2), pp. 249–268.

Hunston, S. (2008) 'Starting with the small words: Patterns, lexis and semantic sequences', *International Journal of Corpus Linguistics*, 13(3), pp. 271–295.

Hunston, S. (2009) 'The usefulness of corpus-based descriptions of English for learners', in *Corpora and language teaching*, p. 141.

Hunston, S. (2011) *Corpus Approaches to Evaluation: Phraseology and Evaluative Language*. New York: Routledge.

Hunston, S. and Francis, G. (1998) 'Verbs Observed: A Corpus-driven Pedagogic Grammar', *Applied Linguistics*, 19(1), pp. 45–72.

Hunston, S. and Francis, G. (2000) *Pattern Grammar: A Corpus-Driven Approach to the Lexical Grammar of English*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Hyland, K. (1996) 'Talking to the Academy: Forms of Hedging in Science Research Articles', *Written Communication1*, 13(2), pp. 251–281.

Hyon, S. (2011) 'Evaluation in Tenure and Promotion letters: Constructing Faculty as Communicators, Stars, and Workers', *Applied Linguistics*, 32(4), pp. 389–407.

Kennedy, G. (2008) 'Phraseology and language pedagogy: Semantic preference associated with

English verbs in the British National Corpus', in Meunier, F. and Granger, S. (eds) *Phraseology in Foreign Language Learning and Teaching*. Amsterdam/Philadelphia: John Benjamins Publishing Company, pp. 21–41.

Klotz, M. (1997) 'Ein Valenzwörterbuch englischer Verben, Adjektive und Substantive', *Zeitschrift für Angewandte Linguistik*, pp. 93–111.

Leech, G. (1974) *Semantics*. Harmondsworth: Penguin.

Li, S. (2015) *A corpus-based study of the high frequency nouns 'time' and 'thing': Investigating the role of phraseology in the construction of meaning in discourse, Department of English Language and Applied Linguistics*. PhD thesis held by the University of Birmingham.

Lindemann, S. and Mauranen, A. (2001) '"It's just real messy": the occurrence and function of just in a corpus of academic speech', *English for Specific Purposes*, 20, pp. 459–475.

Louw, B. (1993) 'Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies', in Baker, M., Francis, G., and Tognini-Bonelli, E. (eds) *Text and Technology: In Honour of John Sinclair*. Amsterdam and Philadelphia: John Benjamins Publishing Company, pp. 157–176.

Louw, B. (2000) 'Contextual prosodic theory: Bringing semantic prosodies to life', in Heffer, C. and Sauntson, H. (eds) *Words in Context: A Tribute to John Sinclair on his Retirement*. Birmingham: University of Birmingham, pp. 48–94.

Louw, B. and Chateau, C. (2010) 'Semantic prosody for the 21st century: Are prosodies smoothed in academic context? A contextual prosodic theoretical perspective', in *Statistical Analysis of Textual Data: Proceeedings of the tenth JADT Conference*, pp. 754–764.

Mahlberg, M. (2005) *English General Nouns: A corpus theoretical approach*. Amsterdam and Philadelphia: John Benjamins.

Malcolm, L. (1987) 'What rules govern tense usage in scientific articles?', *English for Specific Purposes*, 6(1), pp. 31–43.

Mason, O. (1997) 'The Weight of Words: An investigation of lexical gravity', in *PALC*, pp. 97–111.

Mason, O. (2000) 'Parameters of Collocation', in Kirk, J. M. (ed.) *Corpora Galore*. Amsterdam: Rodopi, pp. 267–280.

Mason, O. (2006) *The Automatic Extraction of Linguistic Information from Text Corpora*.

McEnery, T. and Hardie, A. (2012) *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.

McEnery, T., Xiao, R. and Tono, Y. (2006) *Corpus-Based Language Studies: An advanced resource book*. New York and Oxford: Routledge.

Mel'čuk, I. (1995) 'Phrasemes in language and phraseology in linguistics', in Everaert, M. et

al. (eds) *Idioms: Structural and Psychological Perspectives*. Lawrence Erlbaum Associates, pp. 167–232.

Mel'čuk, I. (1998) 'Collocations and Lexical Functions', in Cowie, A. P. (ed.) *Phraseology: Theory, Analysis, and Applications*. Oxford: Clarendon Press.

Morley, J. and Partington, A. (2009) 'A few Frequently Asked Questions about semantic — or evaluative — prosody', *International Journal of Corpus Linguistics*, 14(2), pp. 139–158.

Nesselhauf, N. (2004) 'What are collocations?', in Nesselhauf, N. and Skandera, P. (eds) *Phraseological Units: Basic concepts and their application*. Basel: Schwabe Verlag, pp. 1–21.

O'Keeffe, A., McCarthy, M. and Carter, R. (2007) *From Corpus To Classroom*. Cambridge: Cambridge University Press.

Paltridge, B. (1994) 'Genre Analysis and the Identification of Textual Boundaries', *Applied Linguistics*, 15(3), pp. 288–299.

Partington, A. (1998) *Patterns and Meanings: Using Corpora for English Language Research and Teaching*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Partington, A. (2004) '"Utterly content in each other's company": Semantic prosody and semantic preference', *International Journal of Corpus Linguistics*, 9(1), pp. 131–156.

Partington, A. (2014) 'Mind the gaps: The role of corpus linguistics in researching absences', *International Journal of Corpus Linguistics*, 19(1), pp. 118–146.

Pawley, A. and Syder, F. H. (1983) 'Two puzzles for linguistic theory: nativelike selection and nativelike fluency', *Language and communication*, pp. 191–226.

Renouf, A. (1987) 'Corpus Development', in Sinclair, J. (ed.) *Looking Up: An account of the COBUILD Project in lexical computing*. London: Harper Collins Publishers.

Renouf, A. and Sinclair, J. (1991) 'Collocational frameworks in English', in Aijmer, K. and Altenberg, B. (eds) *English Corpus Linguistics: Studies in Honour of Jan Svaartvik*. New York and London: Routledge.

Sinclair, J. (1987) 'The Nature of the Evidence', in Sinclair, J. M. (ed.) *Looking Up: An account of the COBUILD Project in lexical computing*. London: Harper Collins Publishers, pp. 150–159.

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

Sinclair, J. (2003) *Reading Concordances: An Introduction*. London: Longman.

Sinclair, J. (2004a) 'New evidence, new priorities, new attitudes', in Sinclair, J. (ed.) *How to use Corpora in Language Teaching*. Amsterdam/Philadelphia: John Benjamins Publishing Company, pp. 271–299.

Sinclair, J. (2004b) 'The lexical item', in Sinclair, J. and Carter, R. (eds) *Trust the Text:*

Language, corpus and discourse. London: Routledge, pp. 131–148.

Sinclair, J. (2004c) 'The search for units of meaning', in Sinclair, J. and Carter, R. (eds) *Trust the Text: Language, corpus and discourse*. London: Routledge, pp. 24–48.

Sinclair, J. (2004d) *Trust the Text: Language, corpus and discourse*. Edited by R. Carter and J. Sinclair. London: Routledge.

Sinclair, J. (2008) 'The phrase, the whole phrase, and nothing but the phrase', in Granger, S. and Meunier, F. (eds) *Phraseology: An interdisciplinary perspective*. Amsterdam/Philadelphia: John Benjamins Publishing Company, pp. 407–410.

Sinclair, J., Jones, S. and Daley, R. (2004) *English Collocation Studies: The OSTI Report*. Edited by R. Krishnamurthy. London and New York: Continuum.

Stefanowitsch, A. and Gries, S. T. (2003) 'Collostructions: Investigating the interaction of words and constructions', *International Journal of Corpus Linguistics*, 8(2), pp. 209–243.

Stewart, D. (2010) *Semantic Prosody: A Critical Evaluation*. New York: Routledge.

Stubbs, M. (1995) 'Collocations and semantic profiles: On the cause of the trouble with quantitative studies', *Functions of Language*, 2(1).

Stubbs, M. (1996) *Text and Corpus Analysis*. Oxford: Blackwell Publishers Ltd.

Stubbs, M. (2001a) 'On inference theories and code theories: Corpus evidence for semantic schemas', *Text*, 21(3), pp. 437–65.

Stubbs, M. (2001b) 'Texts, corpora, and problems of interpretation: a response to Widdowson', *Applied Linguistics*, 22(2), pp. 149–172.

Stubbs, M. (2001c) *Words and phrases: Corpus studies of lexical semantics*. Oxford: Blackwell Publishing.

Stubbs, M. (2002) 'Two quantitative methods of studying phraseology in English', *International Journal of Corpus Linguistics*, 7(2), pp. 215–244.

Stubbs, M. (2007) 'Quantitative data on multi-word sequences in English: the case of the word world', in Hoey, M. et al. (eds) *Text, Discourse and Corpora: Theory and Analysis*. London: Continuum, pp. 163–189.

Stubbs, M. and Barth, I. (2003) 'Using recurrent phrases as text-type discriminators: A quantitative method and some findings', *Functions of Language*, 10(1), pp. 61–104.

Summers, D. (1996) 'Computer lexicography: the importance of representativeness in relation to frequency', in Thomas, J. and Short, M. (eds) *Using Corpora for Language Research*. London: Longman, pp. 260–266.

Swales, J. (1990) *Genre Analysis: English in academic and research settings*. Cambridge: Cambridge University Press.

Swales, J. *et al.* (1998) 'Consider this: the role of imperatives in scholarly writing', *Applied Linguistics*, 19(1), pp. 97–121.

Swales, J. and Najjar, H. (1987) 'The writing of Research Article Introductions', *Written Communication*, 4(2), pp. 175–191.

Thompson, G. and Yiyun, Y. (1991) 'Evaluation in the reporting verbs used in academic papers', *Applied Linguistics*, 12(4), pp. 365–382.

Thompson, P. (2009) *Literature Reviews in Applied PhD Theses: Evidence and Problems, Academic Evaluation: Review Genres in University Settings*. Edited by K. Hyland and G. Diani. Palgrave MacMillan.

Tognini-Bonelli, E. (2001) *Corpus Linguistics at Work*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Tognini-Bonelli, E. (2002) 'Functionally complete units of meaning across English and Italian: Towards a corpus-driven approach', in Altenberg, B. and Granger, S. (eds) *Lexis in Contrast*. Amsterdam/Philadelphia: John Benjamins Publishing Company, pp. 73–95.

Walker, C. P. (2008) *A corpus-based study of the linguistic features and processes which influence the way collocations are formed: Some implications for the learning of collocations*. Phd Thesis held by the University of Birmingham

Walker, C. P. (2011a) 'A corpus-based study of the linguistic features and processes which influence the way collocations are formed', *TESOL Quarterly*, 45(291–312).

Walker, C. P. (2011b) 'How a corpus-based study of the factors which influence collocation can help in the teaching of business English', *English for Specific Purposes*. Elsevier Ltd, 30(2), pp. 101–112.

Wei, N. and Li, X. (2014) 'Exploring semantic preference and semantic prosody across English and Chinese: Their roles for cross-linguistic equivalence', *Corpus Linguistics and Linguistic Theory*, 10(1), pp. 103–138. doi: 10.1515/cllt-2013-0018.

Weizman, E. (1984) 'Some Register Characteristics of Journalistic Language: Are They Universals?', *Applied Linguistics*, 5(1).

Whitsitt, S. (2005) 'A critique of the concept of semantic prosody', *International Journal of Corpus Linguistics*, 10(3), pp. 283–305.

Wray, A. (2002) *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press.

Xiao, R. and McEnery, T. (2006) 'Collocation, semantic prosody, and near synonymy: A cross-linguistic perspective', *Applied linguistics*, 27(1), pp. 103–129.