Identifying non-compositional idioms in text using WordNet synsets

by

Faye Rochelle Baron

A thesis submitted in conformity with the requirements
for the degree of Master of Science
Graduate Department of Computer Science
University of Toronto

Abstract

Identifying non-compositional idioms in text using WordNet synsets

Faye Rochelle Baron

Master of Science

Graduate Department of Computer Science

University of Toronto

2007

Any natural language processing system that does not have a knowledge of non-compositional idioms and their interpretation will make mistakes. Previous authors have attempted to automatically identify these expressions through the property of non-substitutability: similar words cannot be successfully substituted for words in non-compositional idiom expressions without changing their meaning.

In this study, we use the non-substitutability property of idioms to contrast and expand the ideas of previous works, drawing on WordNet for the attempted substitutions. We attempt to determine the best way to automatically identify idioms through the comparison of algorithms including frequency counts, pointwise mutual information and PMI ranges; the evaluation of the importance of relative word position; and the assessment of the usefulness of syntactic relations. We discover that many of the techniques which we try are not useful for identifying idioms and confirm that non-compositionality doesn't appear to be a necessary or sufficient condition for idiomaticity.

Dedication

To my parents, Peter and Vita Baron, with love

Acknowledgements

The completion of this thesis represents the fulfillment of a dream and the end of a chapter in my life. I would like to express my gratitude to all who have contributed to this rich experience and helped make it possible. I thank Graeme Hirst and Suzanne Stevenson, my supervisors, for their guidance through this process. Without Graeme's knowledge, patience, empathy, compassion and wit, I would never have arrived at this point. Suzanne has given generously of her time, week after week, applying the necessary motivating pressure to drive me to the finish line. Without Suzanne's experienced guidance and strategy for testing, I would have had no way to finalize this research. I would also like to thank Steve Easterbrook for his time and support during the period when I was investigating the unrelated Requirements Engineering area, and for allowing me to do research for him.

My lifemate Bruce has been an infinite source of nurturing. Lisa and Erich have given me an incredible amount of support — providing loving encouragement. Saretta and Ira have always been there — with great suppers and a space to stay downtown. My family has been a great source of strength and encouragement.

While it is not possible to name everyone, I would like to thank Tobi Kral, Stephanie Horn, Amber Wilcox-O'Hearn, Harold Connamacher, Richard Krueger, Marge Coahran, Mehrdad Sabetzadeh, Shiva Nejati, Vivian Tsang, Afra Alashi, Saif Mohammed, Uli Germann, Afsaneh Fazly, Paul Cook, Tim Fowler, Paul Gries, Steve Engels, Danny Heap, and Jen Campbell — the list goes on. These wonderful friends that I have made here at the University of Toronto have shared their warmth and intellect to make this experience fun and interesting.

Dekang Lin provided excellent clarification of his work, and Kathleen McKeown provided the (unsupported) Smadja code to broaden my understanding of his technique. Afsaneh Fazly's word-pair lists and reading suggestions were invaluable to this study.

I sincerely thank all these wonderful people for their contributions to this work.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Defining Idioms

There is not a single definition of idioms, and researchers present contrasting concepts of idiomaticity. These views do not necessarily contradict each other; rather, they may complement each other to create a broad perspective on idioms.

In the following subsections, we present several of these viewpoints, highlighting some key attributes that will be relevant for this thesis.

1.1.1 A spectrum of word combinations

McKeown and Radev (2000) place all word combinations on a “continuum”, from free-word combinations through collocations to rigid-word combinations, that is idioms, with no clear demarcation or boundary between these three categories.

The meaning of a free-word combination is, by definition, composed of the meaning of its words: words are combined freely to create a new meaning. An idiom, on the other hand, is *non-compositional* — its meaning is not composed of the meaning of its words. In the most rigid idioms, the words cannot be varied in any way, as the free word combination words can. *Collocations* lie between these two extremes. While collocative words usually retain their meaning in combination, they tend to recurrently combine in the same way, showing a natural affinity for one another, which resists substitution of synonyms in normal usage. Table 1.1 illustrates the distinction between these three word combination types.

Table 1.1: Examples of each type of word combination from McKeown and Radev (2000)

Free word combinations:	the side of the road, to buy a house
Collocations:	to trade actively, table of contents
Idioms:	to kick the bucket, dead end

end

1.1.2 From pure idioms to open collocations

In the introduction to their idiom dictionary, Cowie et al. (1983) argue that strictly defining idioms as non-compositional phrases excludes many expressions that should be regarded as idiomatic. They classify word combinations as *pure idioms*, *figurative idioms*, *restricted collocations*, and *open collocations* — a classification predicated on the degree of substitutability of the constituents and degree of rigidity of the expression.

Pure idioms are *fixed* word combinations that have been established through continual use over time. They are non-compositional in meaning, and do not permit substitution of words by similar words. Figurative idioms are those phrases that have both a literal and a non-compositional (figurative) meaning. The figurative interpretation is the more common, and the literal one is seldom, if ever, applicable. For example, when we say that someone *kicked the bucket*, we usually mean that the person died. However, in certain contexts, we could intend this to mean that someone literally kicked some bucket. Restricted collocations or *semi-idioms* cross the threshold between idioms and collocations, with a foot in either domain. They tend to be made up of a combination of literally and figuratively interpreted words, and are usually used in a specific context. For example, in the expression *blind alley*, the word *alley* can be interpreted literally and, since an alley does not actually see, *blind* is obviously figurative. Restricted collocations include groups of words that are usually found in combination more often than by chance. Open collocations are analogous to the free-word combinations described by McKeown and Radev (2000). Table 1.2 illustrates these four types.

Table 1.2: Examples of each type of word combination from Cowie et al. (1983)

Pure idioms:	blow the gaff, between a rock and a hard place
Figurative idioms:	catch fire, a narrow shave
Restricted Collocations:	jog someone's memory, a blind alley
Open collocations:	fill the sink, broken window

1.1.3 Moon's criteria for idiomaticity

Moon (1998) describes three primary characteristics of idiomaticity: *institutionalization*, *lexicogrammatical fixedness*, and *non-compositionality*. Institutionalization is the acceptance of an expression as a single lexical item, and usually occurs over time. Since language is fluid, institutionalized expressions that are accepted at one period of time might no longer be accepted at another. Lexicogrammatical fixedness or *formal rigidity* occurs when words in an expression occur only in specific forms, with some restrictions. A commonly cited example of this is *to shoot the breeze*, 'to chat'. As illustrated by Fazly (2007), though the verb in this expression could be present in any inflected forms (*shooting the breeze, shot the breeze, shoot the breeze*), more general syntactic variations (**the breeze was shot, *shoot the breezes, *shoot the gentle breeze*), are not possible with the same meaning.

1.1.4 Semantic compositionality

Contrary to Cowie et al. and McKeown and Radev, Nunberg et al. (1994) suggest that idiomatic phrases are, for the large part, semantically compositional, and seldom completely rigid. Like Moon's lexicogrammatical fixedness, they believe that the words in idioms can be present in more than one form albeit not in every form. They acknowledge that some idioms are inflexibly rigid, but claim that these comprise the smaller portion of all idioms. Since parts of many (but not all) idioms are modifiable using adjectives and relative clauses, these parts must have a well-understood meaning. Thus they are at least partly semantically compositional. We give some examples cited by Nunberg et al. in table 1.3. In these examples, it is claimed that some of the words (which we have emphasized in boldface) must be semantically recognizable in order to be modified. This contrasts with the broadly accepted concept of non-compositionality.

Table 1.3: Examples of partly compositional idioms from Nunberg et al. (1994). If an idiom is modifiable, then the modified word or words must be semantically understood

Idiom	*Modified idiom*
leave no **stone** unturned	leave no legal stone unturned
pulling **strings**	Pat got the job by pulling strings that weren't available to anyone else.
touch a **nerve**	touch a couple of nerves

1.1.5 Gluing concepts together

Clearly, there are varying and possibly not wholly consistent viewpoints on what constitutes an idiom. These viewpoints can be reconciled if we regard them as elements that can be combined to form a more complete picture. In this thesis, however, we will use the simpler characteristics of non-compositionality and rigid construction, as agreed on by Cowie et al. (1983) and McKeown and Radev (2000), as our cues to idiomaticity.

1.2 Motivation and goals

Any natural language processing system that does not have a knowledge of idioms and their non-compositional interpretation will make mistakes. For example, if the system translates the idiomatic expression *to kick the bucket* into French, it could not translate the individual words and expect to communicate the same meaning. Rather, it should say *mourir*, 'to die' or perhaps even better, *casser sa pipe*, 'break his pipe' which is the equivalent French idiom. If a system were to perform a search for information on buckets, it should not expect to retrieve documents containing the idiom *kick the bucket*. To prevent the abuse of idiomatic expressions in natural language processing, they must be identified in a lexicon that they may be given the special treatment they require.

The goal of this study, therefore, is to investigate techniques for identifying non-compositional idioms from natural language text in order to build such a lexicon. We use substitutability tests to exploit the non-compositional characteristic of idioms. For freely combining words, such as *give a present* we can substitute similar words for the components to create an expression with a similar meaning (*give a gift, donate a present*. However, idiomatic expressions fail substitutability tests because their meaning cannot be derived from the meaning of their parts. For example, while one can say *Susan kicked the bucket* and mean that Susan died, one cannot substitute *pail* for *bucket*, creating the expression *Susan kicked the pail*, and still mean that Susan died. We examine the positional relations of co-occurring words to exploit the rigidity property of idioms. We test several hypotheses:

1. Since idioms are non-compositional, when we substitute similar words for words in an idiomatic expression, the newly-formed expression is seldom, if ever found.

2. When testing for compositionality, not only may similar words be substituted for words in an expression, but also antonyms and other related words.

3. Idioms are rigid expressions whose constituents cannot be rearranged. Therefore,

the relative position of the words in an idiom must be maintained.

We also look at three algorithms to measure the compositionality through substitutions. We consider pointwise mutual information (PMI), introduced by Church and Hanks (1989) to measure word association strength. We extend PMI to incorporate a confidence factor, similar to the work of Lin (1999). Our third algorithm is a simple frequency measure which looks at occurrences of word-pairs and their part-of-speech (POS) tags. Such frequencies have been used by both Justeson and Katz (1995) and Pearce (2001).

We use the British National Corpus to develop a model of the English language, against which we test the word-substitution-hypotheses in both idiomatic and non-idiomatic word-pairs, and identify the best algorithm for measuring substitutability.

1.3 Outline of study

The remainder of this thesis is organized as follows:

Chapter 2, Related work: discusses prior approaches to the identification of collocations and non-compositional expressions. It is on this previous research that we build our study.

Chapter 3, Evaluating techniques and measures for extracting idioms: describes the purpose and approach of this study, providing the underlying motivation for what we are doing.

Chapter 4, Materials and methods: gives a detailed step by step of how the study is conducted. In this chapter, we describe the data, our corpus, any data structures used, and all tests that we perform.

Chapter 5, Experimental Results: discusses the outcome of our tests. We analyze the results and present possible explanations which address the reasons for the (unsatisfactory) outcome.

Chapter 6, Conclusions: looks at the contributions made by this study and suggests follow-on work which could possibly improve our ability to understand and extract idioms.

Chapter 2

Related work

In this chapter, several approaches to the automatic identification of idioms, specialized terminology, and domain-specific collocations are discussed. Where possible, we relate linguistic cues to these techniques. Finally, we examine some measures that have been used to differentiate idioms and collocations from free-association expressions.

2.1 Techniques to identify collocations and idioms

Considerable work has been done in the identification of collocations, technical jargon and domain-specific expressions, and idioms. In this section, we look at some of these efforts including: samples of the earliest research in this area; work predicated on the idiomatic property of non-compositionality; identification through obvious language translation mismatches; and the implementation of latent semantic analysis and asymmetric directed graphs to detect idioms.

2.1.1 Early work involving patterns, frequencies and part-of-speech tags

Choueka et al.

Choueka et al. (1983) are responsible for some of the earliest work in collocation identification on the RESPONSA database (consisting of 176 volumes of Rabbinical documents dating back over a thousand years). They derive an algorithm to identify likely candidates which is predicated on word co-occurrence frequencies. It is based on the diversity of neighbours of each word and their variance from the mean co-occurrence.

Smadja

Smadja (1993) uses positional frequency to help determine the likelihood that a word-pair belongs to a domain-specific collocation expression. He looks at the co-occurrence frequency of a specific pair of words occurring in each position up to five words apart. These frequencies are then compared using a z-score based on the mean and standard deviation. When a word-pair occurs more frequently in one position than in the other four other positions — at least one standard deviation above the mean — he selects that word-pair, preserving the relative position between them, as a candidate collocation. This is illustrated in Figure 2.1.

Once all candidate word-pairs and their relative positions have been selected, Smadja then goes back to examine the context in which they occur in the corpus. Each selected word-pair is aligned in the corpus, creating a *concordance* for all occurrences of that word-pair which preserves the selected distance between the words. Figure 2.2 shows what this might look like for the word-pair *son* and *gun* at a distance of two words apart. He then looks at the words occurring in positions between and around the word-pair. If any one word, (or in some cases word-class such as pronoun), occurs more than fifty percent of the time in a position, it is considered to be part of the multi-word phrase in this position. In this manner, Smadja, identifies multi-word phrases in the corpus. Looking at Figure 2.1, we would expect that *of* and *a* would fit this criteria, and the idiom *son of a gun* would be identified.

This work provides critical insight into the usefulness of preserving the positional relationships between word-pairs in rigid phrases.

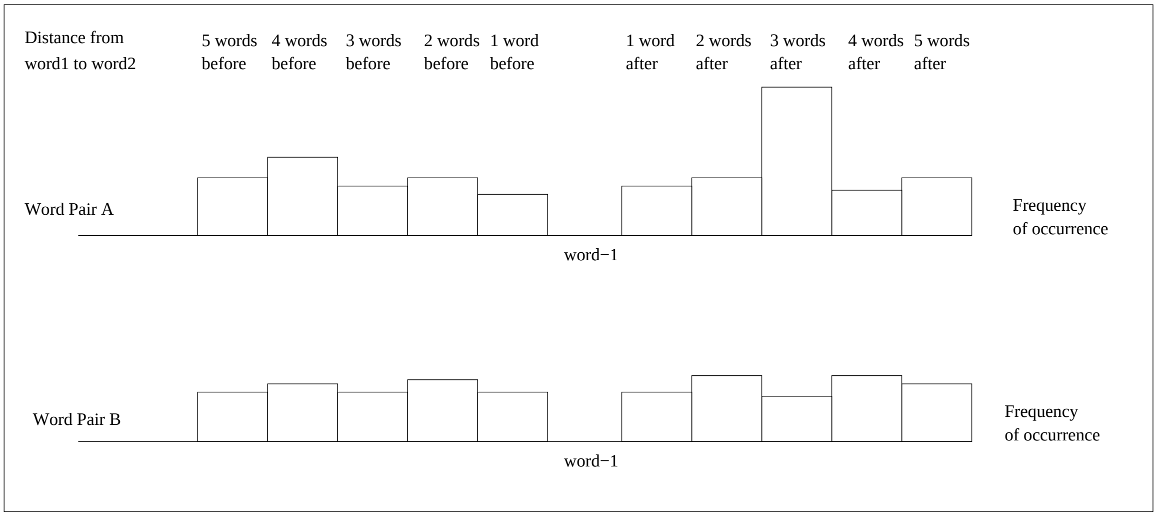

Figure 2.1: If two words co-occur more frequently at one distance from each other than at other distances, as in word-pair A at three words after, that co-occurrence is a more likely collocation candidate. If the co-occurrences at all distances are more or less the same, as in word-pair B, then none of the co-occurrences are candidates.

Justeson and Katz

Justeson and Katz (1995) identify technical terminology, which can be thought of as a kind of idiom, using frequency, part-of-speech tags, and specific orderings of word types, with a precision of 67% to 96%. They do not compare their results against any baseline. Through the examination of medical and other domain-specific documents, they determined that technical jargon consists primarily of noun phrases consisting of nouns and/or adjectives. Occasionally, the preposition *of* is used. As well, technical terms are usually wholly repeated in texts. To truncate or shorten them in any way would reduce

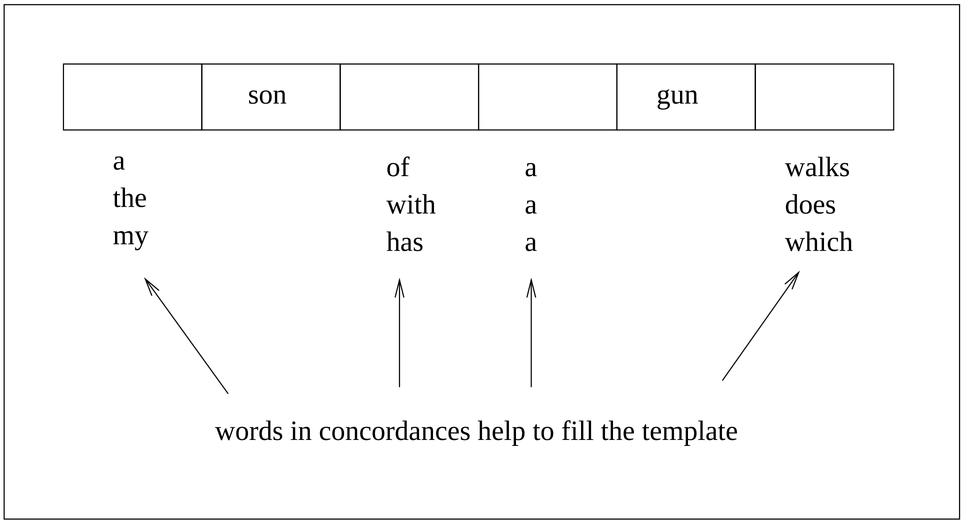

Figure 2.2: Once a word-pair has been selected as a collocation candidate, every occurrence of it in the corpus is extracted and aligned on the word-pair, preserving the positional relation between words, to create a *concordance*.

the information conveyed in the expression, thus changing their meaning.

Justeson and Katz's algorithm is quite simple.

1. Candidate expressions must occur at least twice in a text.

2. Candidate expressions must satisfy the regular expression $((A \mid N)^{+} \mid ((A \mid N)^{*}(NP)^{?})(A \mid N)^{*})N$, where A is an adjective, and N is a noun.

Table 2.1 illustrates the patterns and some of the expressions that were identified in this manner. In order to eliminate the overhead of precisely tagging the words with their part-of-speech (POS), they implement a simplified tagging algorithm. Using a lexicon to identify possible POS tags for a word, they automatically classify it as a noun, adjective or preposition, in that order, if the POS classification for that word exists.

While this work does not identify idioms per se, it illustrates the application of word patterns based on parts of speech as a useful tool for extracting rigid expressions.

Table 2.1: Sample word patterns used for extraction and occurrence examples drawn from three domains, found by Justeson and Katz (1995).

Word Pattern	*Examples*
AN	linear function; lexical ambiguity; mobile phase
NN	regression coefficient; word sense; surface area
AAN	Gaussian random variable; lexical conceptual paradigm; aqueous mobile phase
ANN	cumulative distribution function; lexical ambiguity resolution; accessible surface area
NAN	mean squared error; domain independent set; silica based packing
NNN	class probability function; text analysis system; gradient elution chromatography
NPN	degrees of freedom; energy of adsorption

2.1.2 Substituting similar words to identify non-compositional phrases

Several researchers use the non-compositional property of idioms as a cue to detection. This technique is predicated on the following reasoning: Non-compositional expressions are expressions whose meanings cannot be derived directly from the meanings of the words of which they are comprised. This suggests that if, on the other hand, an expression is compositional, words in the expression can be replaced by words similar in meaning without greatly changing its meaning. Taking this one step further, if an expression is compositional, then it can be assumed that at some point, in a broad sample of the language, such alternative forms of the expression will be observed. Otherwise, the expression is probably not compositional, but rather an idiom. That is, if a word collocates with another word, and a similar word seldom or never collocates with that word, then the collocation probably has a special meaning and strength of association that does not depend solely on the meaning of the individual words. Both Lin (1999) and Pearce (2001) use this kind of substitutability of similar words for identifying non-compositional phrases. Fazly (2007) uses this technique as well as part of her investigation of idioms.

Lin

Lin combines pointwise mutual information ranges (which we will discuss in section 2.2.3) with substitutability in order to identify non-compositional phrases. Using his Minipar parser (Lin, 1998b), he first extracts *triples*, consisting of two words and their syntactic relation, from the corpus. He then calculates the pointwise mutual information (PMI) range for the elements of each triple, incorporating a confidence factor to adjust for the possibility that the frequency of the words in the corpus does not accurately reflect the real world. Then, using his technique for creating a dictionary of similar words (Lin, 1998a), he automatically extracts a thesaurus of similar words from the corpus. For each word in a triple, similar words from this thesaurus are substituted and the mutual information range is calculated for the triples that result. If the range of the original triple is higher than all of the new ones formed, and does not overlap with them, then the original triple is deemed to be non-compositional.

To evaluate his method, Lin selects ten words and manually identifies in a lexicon all idioms which use these words in a pre-specified syntactic format. He then determines which of these idioms occurring in the corpus are identified using PMI ranges. Though Lin achieves only 15.7% precision and 13.7% recall, he points out that that even linguists do not agree completely on idiomatic expressions since a different idiom lexicon scores 39.4% precision and 20.9% recall when classifying expressions from the lexicon used for his study — illustrating a significant difference in opinion.

Pearce

Pearce (2001) follows Lin's (1999) lead, but uses WordNet to identify synsets (groups of synonyms) as words to be substituted in bigrams he has extracted from text. He examines the occurrence frequency of bigrams in two sources: the British National Corpus (2000), and World Wide Web. The frequency of each bigram is compared to the frequencies of those bigrams created through the substitution of words from the synsets for words in the bigram. If the original bigram frequency is much greater than that of the resulting word-pair, the original is considered a collocation. This algorithm is less complicated than Lin's, but, in the sample results Pearce has provided, it seems to perform the task of identifying bigram collocations well.

Fazly

Fazly (2007) provides a more sophisticated and thorough approach to the identification of verb-noun idiomatic combinations (VNIC). Both the non-compositional (lexical fixedness) property of idioms and the syntactic fixedness aspect are explored; the latter directly addresses the claim by Nunberg et al. (1994) and Moon (1998) that most idioms are not strictly rigid, but are often found in a restricted set of modified forms. To study the lexical fixedness of idioms, Fazly uses substitution in verb-noun word-pairs, extracting substitutable words using Lin's (1998a) thesaurus of similar words. To examine syntactic fixedness, she looks at the passivization of the verb, since idioms are not usually present in passive form; at the determiner used; and at the morphology of the noun in the expression, since morphology such as pluralization tends to imply that the noun is more literal and less figurative. These fixedness properties are examined independently and in combination to determine their effectiveness in identifying idioms. The accuracy of syntactic fixedness (71%) as an idiom identification property is slightly better than that of lexical fixedness (68%). The accuracy of the combination of both properties is even better (74%). This is a considerable improvement over the baseline accuracy (63%)

using PMI (see Section 2.2.2). This study clearly demonstrates that the process of idiom identification is improved when other properties are used besides collocation strength and simple substitution.

2.1.3 Other approaches

Here, we look at some other techniques to identify idioms. Melamed (1997) looks at how non-compositional word compounds are translated to other languages; Katz and Giesbrecht (2006) examine the context in which expressions occur; and Widdows and Dorow (2005) use directed asymmetric graphs to find idioms and other closely related word-pairs.

Melamed

Melamed (1997) uses parallel bilingual texts to discover non-compositional compounds (NCCs). His work is premised on two assumptions about NCCs. The first assumption is that the multiple words that make up an NCC in one language are sometimes translated into a single word in another. When this occurs, the meaning of the group of words is derived not from the individual constituents but from the entire group. His next assumption is that in an NCC, at most one of two adjacent words in the *source text* can be linked to the *target text*. The process of discovery uses iterative translations. In each iteration, he applies an estimation function which is an extension of mutual information, to predict whether or not a bigram is an NCC. If it is, he adds it to his NCC list, and for the next iteration, he *fuses* the bigram into a single word which can be successfully linked from source to target. Though he does not report the accuracy of his NCC identification, he does report the improvement in translation using pre-identified NCCs over each translation iteration. While the approach works to some degree, the unrealistic assumption about translation of multiple words to a single word limits its coverage.

Katz and Giesbrecht

Building on Firth's (1957) contextual theory of meaning, predicated on his philosophy that you can tell the meaning of a word by the company it keeps, Katz and Giesbrecht (2006) use latent semantic analysis (LSA) (Deerwester et al., 1990) to differentiate between compositional and non-compositional multi-word expressions. This work differs from the other research we have examined thus far — rather than identify non-compositional expressions, Katz and Giesbrecht examine each expression *in situ* to classify each occurrence as either compositional or not. Whereas idiom identification is a strict either-or categorization of an expression — an expression can either be compositional or non-compositional; their classification technique may classify an expression as non-compositional in one instance and compositional in another, depending on its use. This allows for those idiomatic expressions that are used both figuratively and literally.

The underpinning of Katz and Giesbrecht's work is that when a word is used compositionally, it will usually be in an appropriate context for that word. When a group of words is non-compositional, the context of that group should differ from the usual context of the individual words. They give this example:

1. *Das Kind war beim Baden von einer Luftmatratze ins Wasser gefallen.*
   'The child had fallen into the water from an air mattress while swimming'

2. *Die Enröfnung des Skateparks ist ins Wasser gefallen.*
   'The opening of the skatepark was cancelled'

It is clear in this example that *ins Wasser gefallen* literally means 'fell into the water' in the first sentence, but has the non-compositional meaning of 'cancelled' in the second. Where it has a literal sense, words contextually suitable to water such as *swimming* and *air mattress* are present. The context of *ins Wasser gefallen* is totally different in the second sentence, suggesting that this use of the expression is non-compositional.

To test their hypothesis, for each word group which they are evaluating, as well as the words which that group is composed of, Katz and Giesbrecht build LSA vectors which express the frequency of collocating words — in effect, modeling their context. The cosine similarity is then calculated between the vectors for individual words that make up a group, and the vector for the word group. If they are dissimilar, then we deduce that they are used in different contexts, and that the word group is therefore non-compositional. This is a clear departure from other techniques since it focuses not on the words in a phrase, but on the context in which they occur.

Widdows and Dorow

Widdows and Dorow (2005) have broadened the definition of an idiom to include historic quotations, titles of well-known works, colloquialisms and groups of fixed-noun expressions. They take as idiomatic noun-pairs that are joined by the conjunction *and* only if they occur in one order and not the other. For example, the expression *buttons and bows* would never appear in the corpus as *bows and buttons*, nor would *Porgy and Bess* be present as *Bess and Porgy*. Using conjunctions as verbal cues for word relationships and properties is not a new idea; Hatzivassiloglou and McKeown (1997) use conjunctions (and disjunctions) to identify semantically similar orientations of adjectives.

While Widdows and Dorow identify an informative type of syntactic fixedness for various types of fixed phrases, this work does not appear to be generalizable. The subset of idioms that they identify represents a minuscule portion of the true idioms in the language — even if we constrain ourselves to the rigid side of the collocation spectrum. As well, their broadening of the definition of an idiom to include a variety of rigid noun-pair types is not justified by linguistic or practical motivations. That they also implement a graph paradigm to relate pairs of words, with words as nodes in the graph, is extraneous to the problem of idiom identification though it may be useful for creating a language model similar to WordNet.

2.2 Calculating idiomaticity

In one way or another, all of the methods that use substitution for identifying idioms rely on some kind of a measure along which the original expression and those formed through substitutions may be compared to evaluate the degree of idiomaticity or non-compositionality of the original. In this section, we look at three possible measures: simple frequency counts, pointwise mutual information, and PMI ranges.

2.2.1 Frequency counts

Both Smadja (1993) and Pearce (2001) use co-occurrence frequency as a measure to select bigrams as collocation candidates. Smadja sets an occurrence frequency threshold which must be met by bigrams in order to be considered as candidates. Pearce, on the other hand, does not clarify what the difference between the occurrence frequency of an original word-pair and that of a pair formed through substitution must be in order to classify the original as a collocation. He merely shows that likely collocation word-pairs occur more frequently than unlikely ones. clearly

Co-occurrence frequency is an unsophisticated measure and therefore does not appear to offer much promise as a measure for identifying non-compositional idioms. However, as Manning and Schütze (2000, ch.5) point out, "surprisingly accurate" results are obtained when frequency is augmented with part-of-speech tagging, as shown by Justeson and Katz (1995). While Pearce does not appear to use part-of-speech tags, he does discuss looking at a word and its modifier. Smadja, on the other hand, pays strict attention to not only the words in a bigram but also their part-of-speech tags.

These previous research efforts involving co-occurrence frequency as a measure are somewhat problematic since, for the most part, they do not outline a frequency threshold or a minimum comparative difference to differentiate idioms from non-idioms. It is not clear whether selecting the word-pair which has the highest frequency and ignoring all others is the best way to select an idiom since this could lead to the false identification of very frequent compositional word-pairs as idiomatic or it could fail to catch idiomatic word-pairs which occur less frequently in the corpus.

2.2.2 Pointwise mutual information

Pointwise mutual information (PMI) is a statistic that is often used to measure the strength of association of a word-pair (Church et al., 1991a). It is defined as:

$$I(x;y) \equiv \log_2 \frac{P(x,y)}{P(x)\;P(y)}$$

That is, PMI is the joint probability of x and y occurring together divided by the probability of x and y occurring independently. PMI is calculated as follows, where N is the number of bigrams in the corpus and $|a|$ is the number of times that some word a occurs in the corpus:

$$\begin{aligned} P(x,y) &= \frac{|x \text{ and } y|}{N} \\ P(x) &= \frac{|x|}{N} \\ P(y) &= \frac{|y|}{N} \end{aligned}$$

$$\begin{aligned} I(x;y) &= \log_2 \frac{\frac{|x \text{ and } y|}{N}}{\frac{|x|\times|y|}{N^2}} \\ &= \log_2 \frac{|x \text{ and } y| \times N}{|x| \times |y|} \end{aligned}$$

If the PMI of a word-pair is high, the words are usually strongly associated (Church and Hanks, 1989). PMI assumes normal distribution of words and fails when data is sparse or an incomplete representation of word-pairs is provided due to faulty language source selection. The latter would occur when the corpus does not sufficiently represent the word-pairs being examined (i.e., insufficient coverage). Manning and Schütze (2000, ch.5) indicate that PMI is more useful for proving the null hypothesis — for showing that there is no difference in the strength of association between two word-pairs — than for proving that there is a difference in the strength of association between them. For example, when calculating PMI for two word-pairs, if the value for both pairs is similar, then this measure cannot be used to determine whether or not they are collocation (or idiom) candidates.

2.2.3 PMI ranges

As we have seen in Subsection 2.1.2, Lin (1999) calculates mutual information ranges to identify non-compositional idioms. We now examine the exact details of this calculation. It is based on a triple that consists of a head-word, its modifier, and the triple type. The triple type includes the part of speech of each word and the syntactic relationship between the words. This is illustrated in Table 2.2. Lin computes the pointwise mutual information (Church and Hanks, 1989) of each triple as:

$$\log \frac{|\text{head type modifier}| \times |* \ \text{type} \ *|}{|\text{head type} \ *| \times |* \ \text{type modifier}|}$$

where:

$|$**head type modifier**$|$ is the frequency of occurrence of the entire triple in the corpus.

$|*$ **type** $*|$ is the frequency of all occurrences of the triple type.

$|$**head type** $*|$ is the number of times the head word participates in the relationship type.

$|*$ **type modifier**$|$ is the number of times the modifier participates in the relationship type.

Similar words are then substituted for both the head and the modifier, and their PMI in the triple type is compared with that of the original triple.

Table 2.2: Example of triple generated by Minipar (Lin, 1998b) for the head word *marry* and its modifier *sister* and a similar pair substituting *sibling* for *sister*.

head	type	modifier
marry	verb-complement-noun	*sister*
marry	verb-complement-noun	*sibling*

Lin assumes that the corpus does not necessarily reflect the occurrence of triples in the real world. To mitigate this possible error, he introduces a slight variation in the triple frequency to generate a PMI range. By adding and subtracting some (small) number to this frequency, a range is created within which the true real-world frequency probably exists. This small amount is equal to the square root of the frequency of the triple times the constant associated with a ninety-five percent confidence factor using a standard two-tailed t-test (Moore and McCabe, 1989). Where z_N is this constant, k is the frequency of a triple, and n is the frequency of all triples, the estimated probability of a specific triple range is calculated as:

$$\frac{k \ \pm \ z_N\sqrt{k}}{n}$$

Lin assumes that the other probabilities in the PMI calculation accurately reflect reality, and do not require the small adjustment needed for the triple occurrence probability. The PMI information range is calculated as:

$$\begin{aligned}
\text{lower bound} &= \log \frac{(\ |\text{head type modifier}| - z_N\sqrt{|\text{head type modifier}|}\ ) \times |*\ \text{type}\ *|}{|\text{head type}\ *| \times |*\ \text{type modifier}|} \\
\text{upper bound} &= \log \frac{(\ |\text{head type modifier}| + z_N\sqrt{|\text{head type modifier}|}\ ) \times |*\ \text{type}\ *|}{|\text{head type}\ *| \times |*\ \text{type modifier}|}
\end{aligned}$$

Lin (1999) uses this PMI range to identify non-compositional collocations as follows:

> A collocation α is non-compositional if there does not exist another collocation β such that (a) β is obtained by substituting the head or the modifier with a

similar word and (b) there is an overlap between the 95% confidence interval of the mutual information values of α and β.

If the PMI range of a triple is higher than and does not intersect the PMI ranges of every triple created through substitution, it is considered a non-compositional phrase.

This method offers a technique for differentiating compositional from non-compositional phrases and a clearly defined identification criteria.

Chapter 3

Evaluating techniques and measures for extracting idioms

In this research, we look at methods and measures to identify idioms. We consider as idioms those expressions described as *idioms* by McKeown and Radev (2000), and as *pure* and *figurative idioms* by Cowie et al. (1983). Exploiting the property of non-compositionality, we use substitutability to differentiate this subset of idioms from other word combinations. Clearly, not all idioms can be identified in this manner, since many idioms are partly compositional or have literal interpretations. Our investigation focuses on empirically determining the best technique for identifying these non-compositional idioms.

We explore techniques to differentiate non-compositional (idiomatic) word-pairs from compositional (non-idiomatic) ones. We focus our investigation on three areas: the importance of preserving the relative position of word occurrence in the corpus; the types of words that can effectively be substituted, and the algorithms used to measure substitutability and thereby determine idiomaticity.

3.1 Relative word position

The earlier works of Smadja (1993) and Justeson and Katz (1995) (discussed in Section 2.1.1) rely on the preservation of both the part-of-speech type and the relative position of two words, when considering their candidacy as part of domain-specific collocations or technical jargon. These types of expression are institutionalized and rigid, disallowing rearrangement of the terms. This suggests that for the subset of idioms which are structurally rigid, the frequency of word-pairs at fixed positions and their POS tags is significant. We investigate the importance of preserving word position in order to identify idioms. Specifically, we compare the correctness of classifications made using the frequency of occurrences of word-pairs anywhere within a distance of five words from each other with the correctness of those that are made using the frequency at a fixed relative position within the five word boundary.

We preserve relative position and POS types when examining alternative word-pairs created through substitution. As well, we count the total co-occurrences of all word-pairs within a distance of five words apart. We refer to this total co-occurrence category as a *bag of words* category since the words could occur anywhere within the five-word boundary. Using the algorithms which we will discuss in Section 3.3, for each word-pair a separate idiomaticity measure is calculated for every co-occurrence up to five words apart. If any one position is determined to be an idiom, then the word-pair is classified as such. A similar set of calculations is performed for all occurrences of the pair within a distance of five words from each other. By measuring the effectiveness of each approach in identifying idioms, we hope to validate the effectiveness of preserving relative word position in co-occurrence frequency.

3.2 Substitutable words

We are using WordNet to provide substitutable words in our tests. Pearce (2001) used WordNet in a similar way, to provide sets of synonyms to be substituted. But synonymy is only one of the word relationships that could be used in substitution to identify compositional expressions. For example, given the expressions:

I hate you.
She played all day.

and substituting antonyms, we create new compositionally comprehensible expressions:

I love you.
She played all night.
She worked all day.

Without further testing we cannot be certain that synonymy is a sufficient criterion for substitutable words. The following identifies other word relationships that we want to investigate so that we may establish their usefulness in discriminating between compositional and non-compositional expressions:

holonym → meronym Find holonyms for each word. Then for each holonym, find the meronyms for that word. For example, starting at *leg* we find *body* as a holonym and take the meronyms of *body* as substitutions for *leg*. So if we had *an arm and a **leg***, we would consider *an arm and a **foot***, or *an arm and a **finger***.

hypernym → hyponym Find hypernyms for each word. Then for each hypernym, find the hyponyms for that word. For example, starting at *red* we find *colour* as a hypernym and take the hyponyms of *colour* as substitutions for *red*. So if we had *a **red** herring*, we would consider *a **blue** herring* and *a **pink** herring*.

antonyms If we had ***pulling** one's leg*, we would consider ***pushing** one's leg*.

By including word sets obtained using these relationships as substitute words in our idiom testing, we can determine whether they should be used as word selection criteria or not.

3.3 Algorithms to measure idiomaticity

While the basic concept of identifying non-compositional expressions through the use of word substitution is straightforward, it is not clear which algorithm should be used, once the word-pairs are formed, to best differentiate idiomatic from non-idiomatic expressions. We compare the effectiveness of the three measures introduced in Section 2.2 — frequency counts, pointwise mutual information, and PMI ranges — to determine the most suitable algorithm to be used. The following subsections adapt these algorithms to our study.

3.3.1 Frequency counts

As discussed in section 2.2.1, though frequency is a rather simple measure, for certain purposes it achieves fairly good results when combined with POS tags. So for each word-pair, including the test pairs and substitution pairs, we record the POS tags and frequency of occurrence of the words at each position up to five words apart. The POS tag of any word substituted must match that of the word in the original word-pair. We also keep the total frequency for all occurrences within a distance of five words (our bag of words). If the occurrence frequency of the test pair is higher than that for all substitution pairs that are a specific distance apart, that pair is classified as idiomatic. We classify the bag of words category using the highest occurrence frequency as well.

3.3.2 Pointwise mutual information

We expand the definition of PMI presented in Section 2.2.2, in a manner similar to that described by Lin (1999), to include three factors in our expression. But whereas Lin's expression uses the syntactic relationship between the words and their POS tags as the type of the triple, our calculation uses the distance between words x and y and their POS tags as type:

$$I(\text{word-1}, \text{type}, \text{word-2}) = \log \frac{|\text{word-1 type word-2}| \times |*\ \text{type}\ *|}{|\text{word-1 type}\ *| \times |*\ \text{type word-2}|}$$

Thus we incorporate and preserve the position factor which we are investigating. The calculation is performed for both test and substitution word-pairs at each distance as well as for the bag of words. If, for any of these discrete (distance) calculation sets, the PMI for the test pair is higher than for all the related substitution pairs, that test pair is classified as idiomatic.

3.3.3 PMI ranges

The third algorithm tested in this research mimics Lin's (1999) PMI range discussed in Section 2.2.3). We modify Lin's algorithm to use the distance between two words instead of the syntactic relationship. We now use the POS tags and the distance as the type:

$$\text{lower bound} \;=\; \log \frac{(\,|\text{word-1 type word-2}| - z_N\sqrt{|\text{word-1 type word-2}|}\,) \times |*\ \text{type}\ *|}{|\text{word-1 type}\ *| \times |*\ \text{type word-2}|}$$

$$\text{upper bound} \;=\; \log \frac{(\,|\text{word-1 type word-2}| + z_N\sqrt{|\text{word-1 type word-2}|}\,) \times |*\ \text{type}\ *|}{|\text{word-1 type}\ *| \times |*\ \text{type word-2}|}$$

If the lowest PMI range of our test word-pair is higher than the highest ranges calculated for all substituted pairs, that pair is considered to be a non-compositional idiom.

For the lower bound to be well defined, we must ensure that the subtraction expression in the numerator evaluates to a number greater than zero, for otherwise the log of the expression cannot be calculated. For this reason, and because we assume a normal distribution, we must restrict this calculation to words which co-occur at least five times in the corpus. When the occurrence frequency is less than the minimum allowed for the PMI range algorithm, the lower range cannot be calculated.

3.3.4 Algorithm limitations

While the PMI range specifies a criterion for classification as an idiom, the frequency count and PMI algorithms have no other criteria than whether or not it has the highest score. Highest score can be problematic when comparing pairs, since one of them must always have the highest score. This would suggest that the highest is always an idiom, which is obviously not the case. Fazly's (2007) PMI z-score (see Section 2.1.2), with a pre-stated threshold, eliminates this "highest value" problem. Smadja (1993) also uses a frequency z-score in his final selection of candidate collocative words. We leave the incorporation of z-scores to establish frequency and PMI thresholds for future efforts.

Chapter 4

Materials and methods

In this chapter we look at the execution details of our research. In order to implement our methods the following procedure is used:

- A model of the language suitable to the goals of our research is created. To accomplish this, we extract information about word co-occurrence in English using the British National Corpus (2000) as a representative sample of the language. (This will be described in Section 4.1.)

- Lists of word-pairs to be classified are created. (This will be described in Section 4.2.)

- Using these lists the occurrence frequencies of the word-pairs are extracted from the corpus. Alternatives to each word in the target pair are then taken from WordNet (Fellbaum, 1998), a lexical resource which links semantically related words. The alternatives are substituted one at a time into word-pairs and the occurrence frequency of the newly-formed word-pairs is extracted. (Section 4.3)

- Different measures, to test for the idiomaticity or compositionality of the original word-pairs, are calculated for each substitution. (Section 4.4)

4.1 Extracting sample language statistics to create a language model

Our language model must contain frequencies of co-occurrence of open-class words within a distance of five words. The POS tags and frequency at each distance apart (one word away, two words away, etc.) must be extracted and stored. We use the British National Corpus (BNC) as a language sample. This section describes the BNC, how it is processed, and how information is kept.

4.1.1 The British National Corpus

The British National Corpus (2000), is composed of 4,054 text samples, each up to 45,000 words in length, extracted from various sources and domains. There are over 100 million words in the corpus. SGML tags provide information about each text, including header tags which give summary information about each. As well, tags are used to break each text into sentences or $< s >$ ... $< /s >$ units and words with their part-of-speech tag or $< w\ POS >$ units. Since we want to look at open-class words, specifically nouns, verbs, adjectives, and adverbs, the tags must be simplified so that a word either fits into one of these simple categories or is excluded entirely. All forms of the verbs *be* and *have* are also excluded as they do not provide semantic information which can be meaningfully substituted. The stop-word list, contained in Appendix A, Table A.1, identifies other words that are deliberately excluded. All headers and extraneous tags are removed. The open-class word tags are simply reclassified as one of N for nouns, V for verbs, J for adjectives, and R for adverbs. Appendix A, Table A.2 illustrates this reclassification.

Processing proceeds sentence by sentence; document structure and document boundaries are completely ignored. Each sentence within a text is individually processed. For each open-class word that exists in the corpus and each subsequent open-class word co-occurring up to five words away, bigrams are created. The POS tags for both words and their distance from each other are also captured with each bigram. As shown in Figure 4.1, starting at the first word and moving toward the last word in the sentence, a window-like template framing the first focus-word and next five words is slid across each word in the sentence.

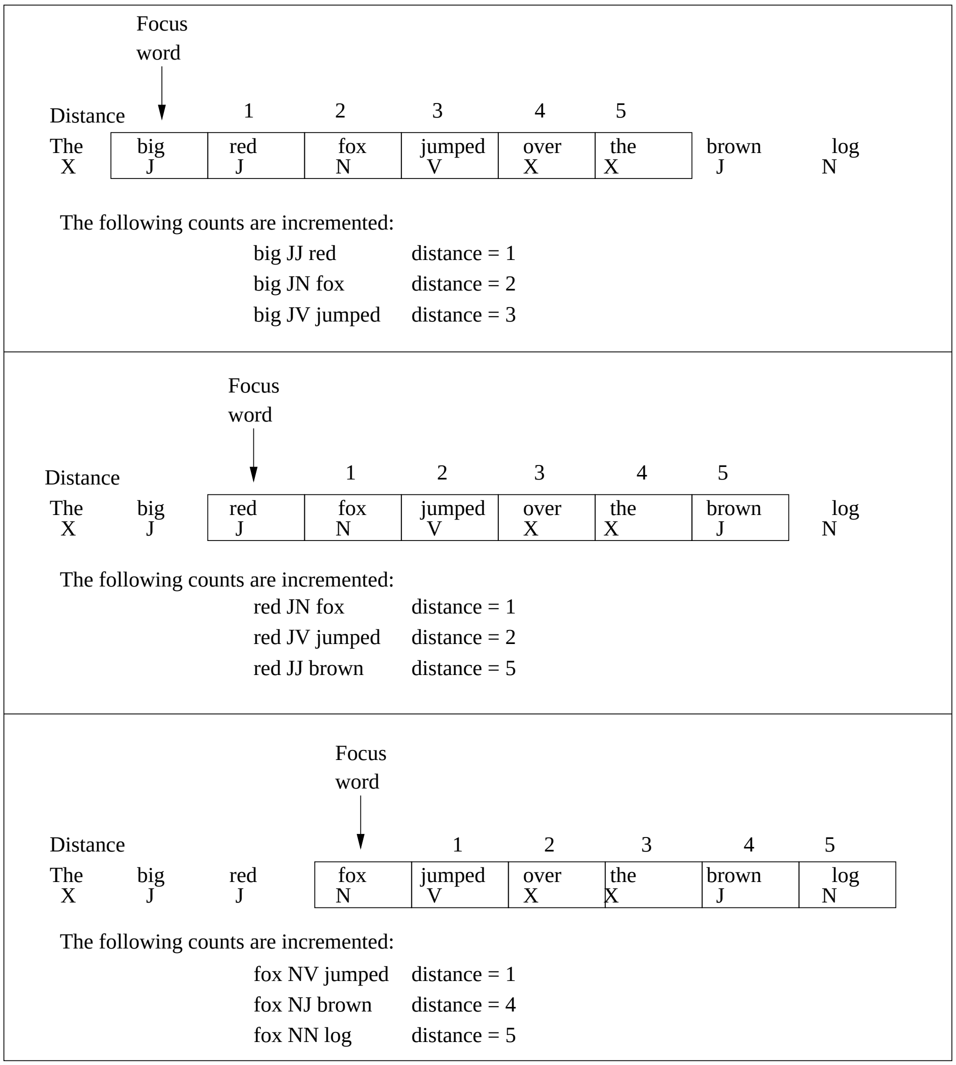

Figure 4.1: For each sentence in the corpus, a sliding window that frames the focus word and subsequent five words, is passed over the words in each sentence. Positional frequency information is collected for open-class word-pairs.

4.1.2 Data structures

It is not possible to process all of the BNC open-class words in memory at one time. For this reason, multiple passes are made through the BNC. Each word being counted and its POS is kept in a hash. Then, for each open-class word type with which it co-occurs, the word, its POS, and the number of times it occurs in each position up to a distance of five words away, is counted. This information is kept in an augmented dictionary abstract data type (ADT) as illustrated in Figure 4.2. When the entire corpus has been examined for co-occurrence of words in the chunk, the counts and other relevant information is stored in a sequential file.

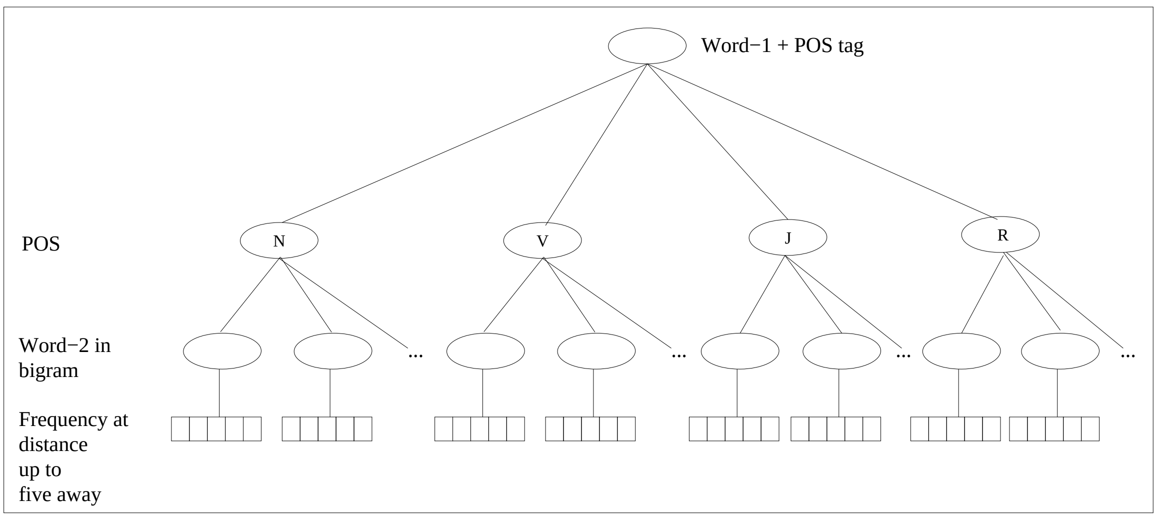

Figure 4.2: Open-class words that are being processed are maintained in a hash in memory. The four open-class POS types are branches for each word, and for each part of speech, the co-occurring words and their occurrence frequency counts are stored.

This information, kept for every word-pair, is treated as a *triple*. As we discussed in Section 3.3.2, our triples contain the first or focus word, the second word, and the type which consists of the word POS tags and distance between the words. Rather than store five different records for each co-occurrence position, this information is maintained in a simple array showing counts for each of the five positions as well as the total co-occurrence counts within a distance of five words. Table 4.1 provides examples of some triples extracted from the corpus.

Table 4.1: For two co-occurring words, the part-of-speech (POS) tags, and co-occurrence frequency counts are maintained. Counts are kept for occurrence in each position up to five words away as well as the total occurrence within five words.

			occurrence counts					
Word-1	*POS*	*Word-2*	*total*	*1 away*	*2 away*	*3 away*	*4 away*	*5 away*
future	NR	past	3	0	1	0	2	0
future	NR	actually	7	1	3	1	1	1
future	NR	only	47	8	18	10	2	9
future	NJ	european	31	0	7	18	3	3
future	NJ	vital	4	0	0	0	4	0
future	NJ	great	23	0	4	14	2	3
future	NN	miners	2	0	0	1	0	1
future	NN	earth	8	0	2	4	2	0
future	NN	railway	10	0	2	4	1	3
future	NV	seems	12	7	3	0	2	0
future	NV	exists	1	0	0	0	1	0
future	NV	lay	37	31	1	1	2	2

In order to compute the PMI and PMI ranges using the algorithms described in Sections 3.3.2 and 3.3.3, we must have frequency counts for the following, where Type is composed of the POS tags for Word-1 and Word-2 and the distance between the words for a specific frequency count:

1. **Word-1 + Type + Word-2:** the number of times the exact triple containing the two words and specified relationship occurs in the corpus.

2. **Word-1 + Type + Any word:** the number of times the first word and specified

relationship occurs with any word.

3. **Any word + Type + Any word:** the number of times the relationship occurs in the corpus with any words.

4. **Any word + Type + Word-2:** the number of times the specified relationship and second word occurs with any first word.

These counts are calculated after all of the triples for all of the open-class corpus words have been extracted to a file. They are maintained in a database for ease of access. Additionally, a data store is created which links the base form of all corpus words to the expanded word form as presented in the corpus. This is discussed in the next section. Figure 4.3 shows the data stores required.

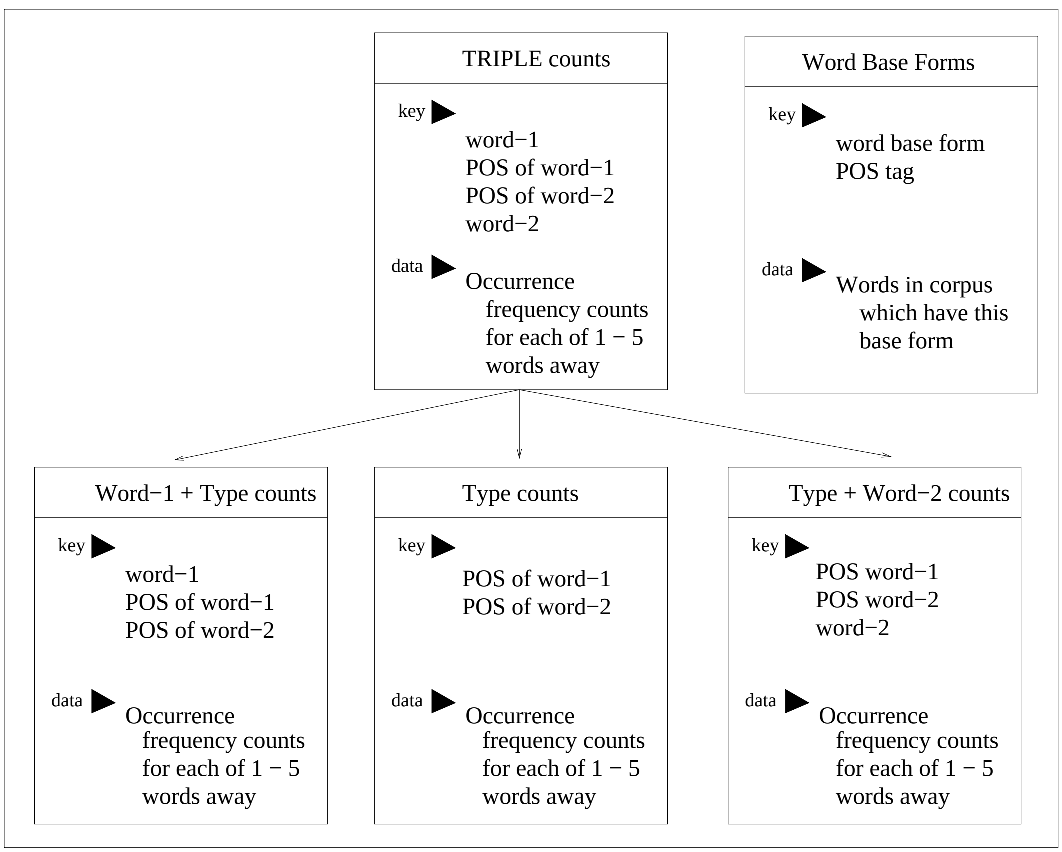

Figure 4.3: Data stores including all necessary fields.

4.1.3 Linking multiple forms to a single base form

The words which we attempt to substitute into our bigrams are provided by WordNet in a stemmed, base form. For example, *burn* as a verb may be present in the corpus as *burned*, *burns*, *burning*, *burnt*, and *burn*, but WordNet would give us only *burn#v*. To ensure that we identify counts for all occurrences of a word, regardless of its form, we must be able to take this base form, and generate keys to access all data using forms of this word.

This is accomplished through a reverse lookup table. The reverse lookup table matches the base form of the word plus the POS tag to a list of all forms of the word for that POS. For example, the entry *burn#v* in the table contains all of the valid forms of *burn* as it is used as a verb in the corpus. We would then substitute each of these forms to get a total occurrence count for the verb *burn*.

4.2 Test data

To test the various idiom recognition techniques, we use lists of word-pairs: Each pair in a list is either part of an idiomatic phrase, or part of a regular compositional phrase. We have three lists, one for development and two for testing. The lists, including corpus occurrence statistics, are available in Section A.2. The first list is used for development: to test concepts, optimize performance, and debug the code. The unseen second and third lists are used for testing.

Two of the lists have been provided by Fazly, and were used in the research for her PhD Thesis (Fazly, 2007). Fazly has carefully vetted her lists with users, using chi-square tests to measure agreement on word-pair classifications. However, not all word-pairs from Fazly's lists could be used, since some of the pairs involve the words *have* and *get*,and hence are not relevant to our study.

Since Fazly's work is primarily concerned with identifying multi-word expressions using light verb and nouns, and our work is not, a third test list (Cowie data) was also constructed by extracting idioms at random from the *Oxford Dictionary of Current Idiomatic English* (Cowie et al., 1983). To create a balance between idioms and non-idioms, for every pair of words in this list, we created a non-idiomatic pair. We paired the first word of the idiom with a free-association word to create a compositional expression (or non-idiomatic pair). Due to time and resource constraints, this list was not validated with users. Fazly's lists have been more rigorously refined and may be reused by others as a *gold standard*; this ad-hoc list should not be.

4.3 Using WordNet for substitution words

WordNet (Fellbaum, 1998), a lexicon which links words by semantic relationships, is used to supply alternative words to be substituted into word-pairs. In addition to synonyms, where possible, we explore other word relation types that may provide substitutable words, as described in Section 3.2, including antonyms, holonym → meronyms, and hypernym → hyponyms. In fact, we run separate trials involving several permutations of relationship types including:

1. synonyms only
2. synonyms and antonyms
3. synonyms, antonyms, and holonym → meronyms
4. synonyms, antonyms, holonym → meronyms, and hypernym → hyponyms
5. synonyms, antonyms, and hypernym → hyponyms.

Using the Perl package WordNet-QueryData-1.45, available through CPAN, as an interface to the WordNet database, we first translate the word to be substituted into its base form. We make no attempt at disambiguation. We search for all senses of this word, and for every sense, find the synonyms, antonyms and other word relations, as necessary. Finally, using the words obtained from WordNet, we search our reverse-lookup table to convert each word from its base form to the forms present in the corpus, stored in our triple database. We substitute each corpus-based word into our triple, in the place of the word we originally searched on, and extract frequency counts for the new, substituted pair.

Where multiple forms of a word are present in triples, all forms are summed into a single frequency count. For example, given the word-pair *drive vehicles*, we would obtain the synsets for *drive* from WordNet. One of these synsets includes the verb *take*. Accessing our reverse-lookup table, we would identify all forms of the verb *take* that are present in the corpus (i.e., *take, took, taken, taking,* and *takes*) and substitute them for *drive* to create new word-pairs. The frequency counts for these pairs would be accrued as though they were a single triple.

Though WordNet contains about 150,000 words, it is limited in size and not available for other languages. This limits our technique to English and languages with a WordNet-like lexicon, and precludes the full automation of this technique. Using a dictionary of automatically extracted related words, as done by Fazly (2007) and Lin (1999), would overcome this barrier and ensure portability of this technique to other languages.

4.4 Calculating idiomaticity

For every word-pair, at each distance of one to five words, and for all occurrences within a distance of five words, we perform the three calculations (discussed in Section 3.3) to determine idiomaticity:

- **Frequency count**: The highest occurrence frequency count for an alternative (substitution) word-pair is subtracted from the occurrence frequency count for the test word-pair.

- **PMI**: We calculate the gap between the PMI of the word-pair and highest PMI score that is obtained by any substituted word-pair.

- **PMI range** The lower-threshold value of the PMI range for our word-pair is calculated. We then calculate the upper-threshold value of the PMI range for every pair obtained through substitution. Finally, we subtract the highest upper-threshold PMI value for all substitutions from the lower-threshold PMI value for the word-pair. (PMI range calculations have been more fully described in Section 3.3.3.)

For each of these calculations, the word-pair is classified as an idiom if and only if the difference is greater than zero. This gives us three separate sets of classifications — one for each calculation.

Chapter 5

Experimental Results

This research focuses on finding the best means to correctly identify non-compositional idioms. To accomplish this, we perform tests to measure three aspects: the importance of maintaining positional co-occurrence frequency counts; the usefulness of additional WordNet relationships; and the relative performance of three selection algorithms. Specifically, we test the classification of word-pairs from lists as either idiomatic or non-idiomatic using substitution — across a full spectrum of permutations of our aspects. We present the empirical outcome of these tests through this chapter. First we define the measures that we will use for comparisons. We then compare the performance of the three measures. Following this, we look at word occurrence frequencies, highlighting the relative importance of preserving frequencies and the relative position in which the words occur when substituting alternative words. Then, the usefulness of augmenting our substitution set with additional words extracted using other WordNet relationships is examined. Finally, we provide an overall view of the results. Additional graphs and tables which show our test results are provided in Appendix B.

5.1 Measuring Results

The classifications assigned by our method are verified against the gold standard label. For each of the three techniques, for all of the WordNet relationship substitution permutations, and for both test lists, we calculate the precision, recall, accuracy and F-score. Precision is the number of word-pairs correctly classified as idioms divided by the total number of word-pairs classified as idioms. Recall is the number of idiomatic word-pairs identified over the total number of idiomatic word-pairs in the test set. Accuracy is the number of pairs classified correctly divided by the total number of pairs. The F-score is calculated as $\frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$. As our baseline, we use the PMI calculation with bag-of-words substitution because it has been used in previous work. Fazly (2007) uses PMI on verb-noun pairs which is not precisely a bag-of-words. However, since her word-pairs are the outcome of parsing, they could be arbitrarily far apart. We interpret this as words which co-occur *somewhere in the neighbourhood of each other* — somewhere in the bag-of-words which make up a sentence. Our bag-of-words is constrained to a distance of five words. The various scores are manually compared, and the best technique for identifying idioms is decided.

It must be noted that throughout the presentation, when we say that some method performs best, unless we are discussing a particular performance measure, we are referring to the overall performance or F-score. While the F-score provides a blend of the precision and recall metrics, using a particular method predicated on this measure is obviously not suitable to all applications — in some instances precision is critical, in others it may be recall. So, whereas a method may outperform another based on the F-score, it may be imprecise and have no practical value. Alternatively, where a method may have an incredibly high precision, it may identify so few idioms that it too is impractical.

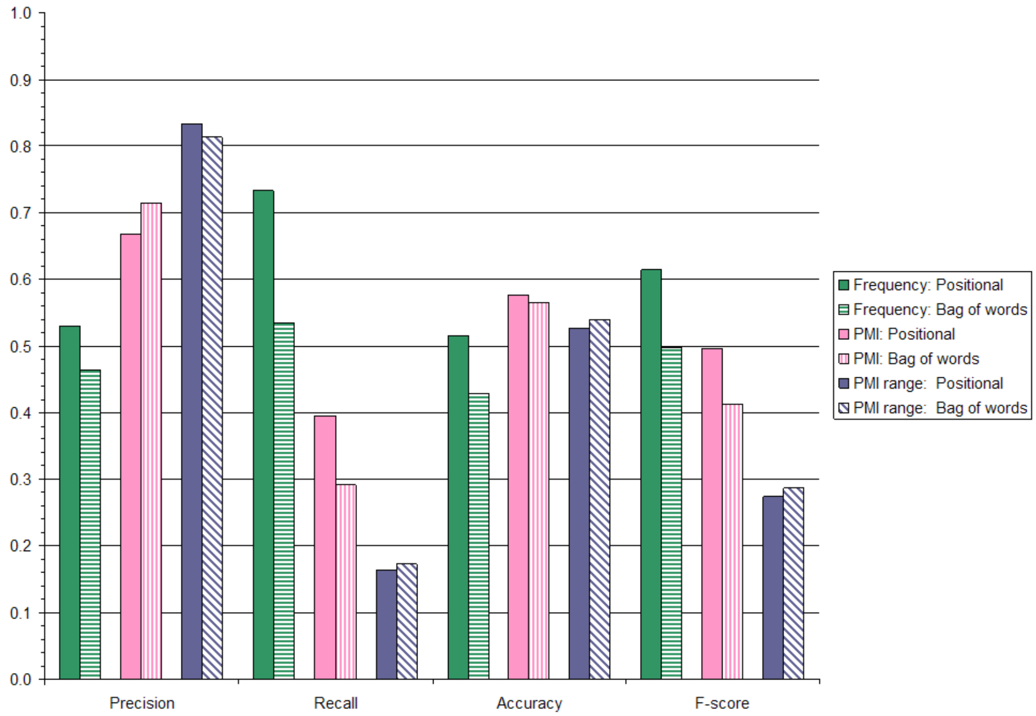

Figure 5.1: The performance of all algorithms when applied to the Fazly test data.

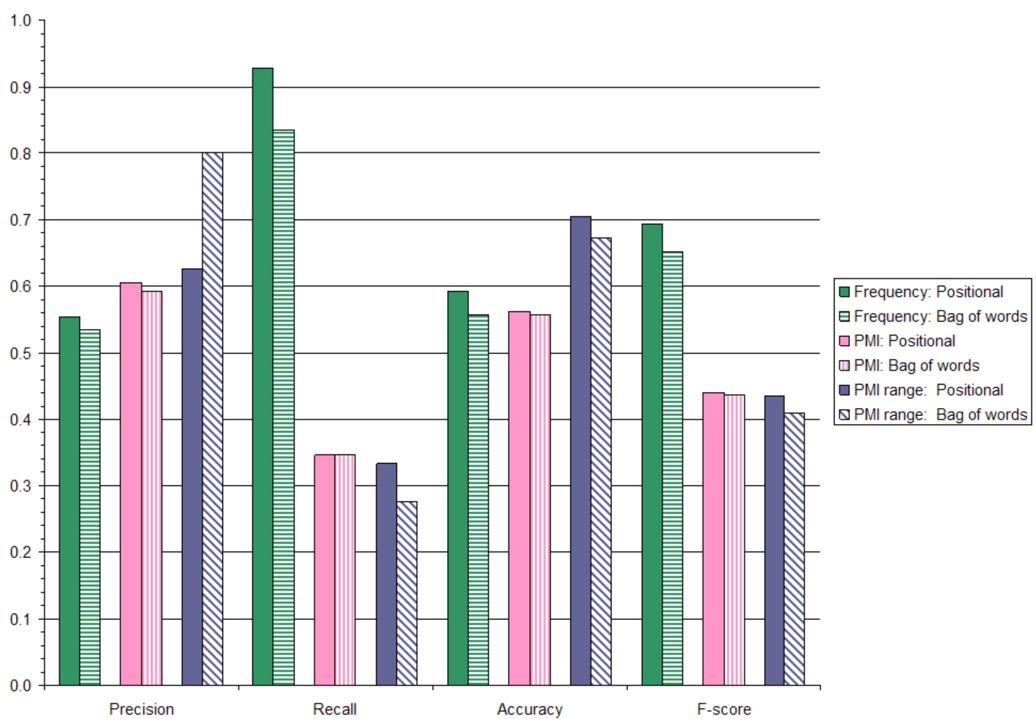

Figure 5.2: The performance of all algorithms when applied to the Cowie test data.

Table 5.1: The results of our tests using both the Fazly test data and the Cowie test data. We show all measures for all algorithms, and constrain our WordNet relationship types to synonyms only.

	idioms found	non-idioms found	Precision	Recall	Accuracy	F-score
			Fazly test data			
Frequency						
Position based	63 of 86	21 of 77	0.53	0.73	0.52	0.61
Bag of words	46 of 86	24 of 77	0.46	0.53	0.43	0.50
PMI						
Position based	34 of 86	60 of 77	0.67	0.40	0.58	0.50
Bag of words	25 of 86	67 of 77	0.71	0.29	0.56	0.41
PMI range						
Position based	10 of 61	49 of 51	0.83	0.16	0.53	0.27
Bag of words	13 of 75	63 of 66	0.81	0.17	0.54	0.29
Average						
Position based			0.68	0.43	0.54	0.46
Bag of words			0.66	0.33	0.51	0.40
			Cowie test data			
Frequency						
Position based	78 of 84	22 of 85	0.55	0.93	0.59	0.69
Bag of words	70 of 84	24 of 85	0.53	0.83	0.56	0.65
PMI						
Position based	29 of 84	66 of 85	0.60	0.35	0.56	0.44
Bag of words	29 of 84	65 of 85	0.59	0.35	0.56	0.44
PMI range						
Position based	5 of 15	26 of 29	0.63	0.33	0.70	0.43
Bag of words	8 of 29	39 of 41	0.80	0.28	0.67	0.41
Average						
Position based			0.5941	0.54	0.62	0.52
Bag of words			0.64	0.48	0.59	0.50

5.2 Algorithm performance

The algorithm performance for the two test data sets are illustrated in Figure 5.1, Figure 5.2, and Table 5.1. For each algorithm we report the results using both the word-pair co-occurrences in each precise word position (positional) and for those which co-occur anywhere within a five word distance (bag-of-words). Our analysis is predicated on the performance comparison between positional and bag-of-word substitutions using synonyms for all three algorithms. We exclude results that incorporate other WordNet relationships since, as we discuss in Section 5.4, these relationships do not seem to significantly contribute to the outcome and cloud our analysis. The results show that the frequency count algorithm, which selects a test pair as an idiom only if the frequency is higher than that for all substituted pairs wins overall as having the highest F-score. However, when we consider precision and recall separately, a different picture emerges.

The PMI range renders better precision. The precision score for the PMI range is 10% and 20% higher than the baseline on the Fazly test data and Cowie test data respectively. However, the algorithm has poor coverage, and it cannot be used where word-pairs occur fewer than five times (Dunning, 1993). As a result, fewer of the word-pairs can be evaluated using this technique — the pair coverage ranges from 26 to 86.5 percent (see table 5.2). So, unless we have a larger corpus than the BNC, the PMI range algorithm, while relatively more precise, is impractical since it cannot be used to evaluate many word-pairs.

As expected, there appears to be a trade off between recall and precision. The frequency algorithm has the highest recall and F-score with values that are on average 51% and 23% higher respectively than the baseline, but in situations where precision is critical, the PMI range algorithm performs best. The PMI and PMI range algorithms are excellent eliminators of non-idioms but they also tend to eliminate many idioms as well. The frequency count algorithm seems to perform in an opposite manner — not only does it classify most idioms as idioms, but also many non-idioms.

When we take a closer look at the individual classifications performed by these algorithms, we see that many assessments using PMI, including the PMI range, because of the deeper word-association measure, eliminate pairs that may occur with high frequency but are not necessarily tightly associated; they may occur with high frequency with other words as well. Unfortunately, because non-compositionality suggests unusual use of a word or words in an expression, the word association measure or PMI value may be too weak to identify a word-pair as compositional when it is.

On the other hand, the frequency algorithm automatically assigns non-compositionality to the word-pair with the highest occurrence count. No consideration is given as to whether those words frequently occur with other words as well. Their association with other words, which is a measure which deepens our understanding of the semantic significance of their relation to each other, is completely ignored. Consequently, while frequency avoids the pitfall of over-elimination that is endemic to PMI, it fails to correctly judge whether or not a word-pair is idiomatic and under-eliminates non-idioms. The idea of using word-pair frequency and POS tags to identify idioms, premised on the work of Justeson and Katz (1995) which uses them to identify specialized terms, does not prove to be fruitful.

We can conclude that for one reason or another, none of these algorithms performs well. It would be interesting to see if they could be synergized into a single algorithm which would incorporate the positive aspects of each part.

5.3 Relative word position

Our tests suggest that it is better to calculate compositionality by preserving position-specific word-pair frequencies than it is to use the frequencies of all occurrences within a five-word distance. Once again, our analysis includes calculations using synonyms only.

As we look at the results presented in Figure 5.1, Figure 5.2, and Table 5.1,we see that calculations using position-specific frequencies of word-pair occurrence have higher precision, recall, accuracy and F-score scores than those which use the bag-of-words occurrence counts including the baseline PMI bag-of-words. Exceptions to this are the precision measure for the PMI calculation on the Fazly test data set and the PMI range calculation on the Cowie data set. The recall measures for both of the bag-of-word calculations are significantly lower. The precision for the bag-of-words PMI range is skewed considerably higher — however, this statistic is misleading, since it evaluates less than half the idioms.

5.4 Alternative WordNet relationships

In addition to synonyms, we used other WordNet relationships to find suitable words for substitution in our tests for idiomaticity (see Section 4.3). We found this not to be useful in any way. We provide the average case results in Table 5.3, and additional charts in Section B.2 which illustrate our performance indicators: precision, recall, accuracy and F-score. In all cases the addition of antonyms performs exactly the same as using synonyms only. Even worse, the recall, accuracy and F-score values degrade when we add any combination of the holonym → meronym or hypernym → hyponym relationships, though in some cases, precision is improved (see Figure 5.3).

We suggest that the reason for this poor performance is that we have over-expanded our substitutable word set. Recall that we use all WordNet synsets for the word to be replaced through substitution (Section 4.3) By contrast, Pearce (2001) does not use a word sense unless he encounters a substitution using at least two different words from that sense in the corpus. By expanding across all senses of a word, as we do, we probably generate too many words and increase the likelihood of finding some in the corpus, false positives, thus wrongly suggesting that the word-pair is compositional. For example, the word-pairs *blow bridge* and *cut cord* occurring seven and ten times respectively, are classified as idioms, having no significant word-pairs found in the corpus using the set of substitutable synonyms from WordNet. However, when the hypernym → hyponym relationship is added, these word-pairs are classified as non-idioms, as the pairs *blow head* and *cut wire* are found in the corpus 14 times and 12 times respectively. For this reason, as we add WordNet relationships to find substitutable words, we find fewer idioms. As we reduce our set of classified idioms, since we have explored a much wider set of substitutable words using all possible relationships, these remaining word-pairs are more likely to be accurately identified. Consequently, while we may improve precision, we significantly reduce recall.

Table 5.2: Coverage of the PMI range algorithm.

	Fazly test data		*Cowie test data*	
	Bag of words	Positional frequency	Bag of words	Positional frequency
Number of eligible idioms	75	61	29	15
Number of eligible non-idioms	66	51	41	29
Actual number of idioms	86	86	84	84
Actual number of non-idioms	77	77	85	85
Percent coverage	87	69	41	26

Table 5.3: The results from word substitution by different WordNet relationships. The results are averaged across all algorithms for both positional and bag-of-words application. The baseline used is the PMI algorithm using bag-of-words substitution. S = synonyms only; A = antonyms; M = holonym → meronym; and H = hypernym → hyponym.

Fazly test data				
	Precision	Recall	Accuracy	F-score
S	0.59	0.44	0.52	0.48
SA	0.59	0.44	0.52	0.48
SAM	0.61	0.42	0.53	0.47
SAH	0.83	0.19	0.53	0.28
SAMH	0.83	0.19	0.53	0.28
Baseline	0.71	0.29	0.56	0.41
Cowie test data				
S	0.58	0.67	0.59	0.60
SA	0.58	0.67	0.59	0.60
SAM	0.58	0.65	0.59	0.59
SAH	0.60	0.46	0.57	0.50
SAMH	0.60	0.46	0.57	0.50
Baseline	0.59	0.35	0.56	0.44

5.5 General analysis

None of the methods we looked at have performed very well. We suggest a number of reasons why they fail:

1. *WordNet limitations*: While WordNet provides an excellent network of semantic information about words, it is at once too broad and too narrow a resource for this purpose. It is too broad, as it provides us with sets of words totally unrelated to the sense of the word in many word-pairs. We provide examples of this in Table 5.4. It is too narrow as it does not contain all of the words for which we are seeking alternatives.

2. *Corpus limitations*: There is a distinct possibility that the corpus does not fairly represent the idiomatic pairs being evaluated. While we cannot directly show evidence of this problem, it could be further validated through the use of a larger corpus such as the 5-grams available from Google (Brants and Franz, 2006) which could be used as pseudo-sliding windows.

3. *Substitutability limitations*: Substitutability is an inadequate criterion for distinguishing non-compositional idioms from compositional expressions. An inability to substitute a similar terms does not necessarily mean that a word-pair is idiomatic. It is possible that the words just tend to collocate more than other similar words. Rather than being a measure of idiomaticity, it is perhaps a better illustration that we tend to select certain words together more than others. For example, we tend to say *fresh ingredients*, but probably would not say *fresh constituents* or *new ingredients*. There are words that we habitually combine the same way but this does not make them idiomatic, merely collocations (Church et al., 1991b).

4. *Data set limitations*: The Fazly data-set consists of light verbs plus nouns. The light verbs do not offer much in the way of semantic information. As a result, any

attempt to substitute synonyms for them is not especially useful. For example the verbs *make*, *get*, and *give* can be combined with almost any of a large number of nouns because so many nouns denote things that can be made gotten or given. Their lack of semantic significance sometimes reduces the value of a word-pair evaluation involving light verbs to a simple noun substitution.

5. *Idiom limitations*: Many idiomatic expressions have literal interpretations which are used as frequently as their figurative ones. Some of the word-pairs which were extracted from an idiom dictionary and classified as idiomatic failed to be identified as non-compositional idioms. Since these word-pairs were used literally as often as they were used figuratively, they were not useful test items. For example, the word-pairs *see daylight*, *cut cord*, *move house*, *cut cloth*, *pull finger*, *give slip*, *see sight*, and *make pile*, which are classified as idiomatic, all appear to be compositional and more non-idiomatic than idiomatic. This problem is eliminated when individual *in situ* classifications are made (Katz and Giesbrecht, 2006).

Our methods do not seem to fail more in one area than another. For one data set, PMI range bag-of-words evaluations are more precise than position-based ones. For the other data set, they are not. This is true of PMI bag-of-word evaluations as well. In one situation, augmenting relations improves performance, in most others, it does not. This lack of consistent performance makes it extremely difficult to identify any single cause of failure.

Table 5.4: The following words were inappropriately substituted in idiomatic word-pairs. They were in fact from an unrelated word sense. As a result, the word-pairs were incorrectly classified as non-idioms. The **boldface** word is the word that is replaced.

Word-1	*Word-2*	*Replacement word*
take	**air**	line
set	**cap** (meaning hat)	ceiling
take	powder	make
see	**red**	loss
find	**tongue**	knife
give	**flick**	picture

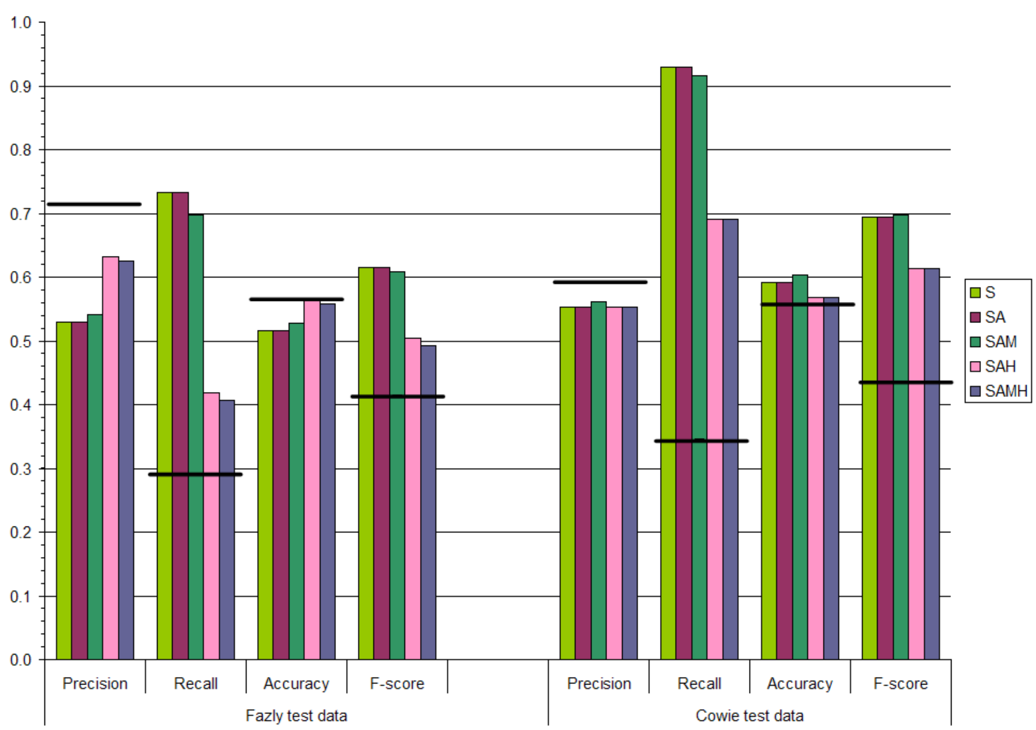

Figure 5.3: The performance of all relationship substitution permutations for both data sets. Including only results for positional frequency using the frequency algorithm. Where S = synonyms only; A = antonyms; M = holonym → meronym; and H = hypernym → hyponym. The baseline, displayed as a black horizontal line, shows the results for synonyms only using the bag-of-words occurrence counts and the PMI algorithm.

Chapter 6

Conclusions

Non-compositional idiomatic expressions pose a significant problem in computational linguistics. Translation, generation, and comprehension of text is confounded by these expressions, since their meaning cannot be derived from their constituent words. Previous research has suggested several techniques for their identification. We have combined and contrasted some of these techniques in an attempt to discover the best way to extract idioms from natural language text. The basic premise, upon which our efforts are built, is the concept that words in these expressions are uniquely combined in a way that does not express their actual meaning and that the expression loses its meaning if similar words are substituted for words in the expression. In fact, by this premise it follows that for any non-compositional idiom, we would never (or rarely) find these substituted expressions in the language.

We have processed the British National Corpus (2000) to create a data model which would permit us to test our ideas. Using two data sets of word-pairs, we looked at the occurrence frequencies of the word-pairs as well as those of pairs formed through the substitution of similar words. The benefit of preserving the relative position of word-pair occurrence over looking at the bag-of-word frequencies, across a five-word distance, has been examined. We have contrasted the performance of three measures: frequency, PMI, and PMI range. Finally, we have measured any improvement gained through augmentation of the WordNet relations from simple synonyms as proposed by Pearce (2001) to include other WordNet relations.

6.1 Summary of contributions

Preservation of word position. Word substitutions are performed using all words in a five-word distance or preserving the relative position of words in each word-pair such that all substitution pairs are the same distance apart as the original test pair. We have shown that, probably because of the pseudo-rigid nature of idioms, substitutions which maintain the original relative word positions do a better job of idiom recognition.

Calculations to identify idioms. We contrast three algorithms that use substitution to identify idioms: comparison of simple occurrence frequency using POS tags; pointwise mutual information; and a PMI range which introduces a confidence factor. Using the PMI bag-of-words as a baseline, we see that though the PMI range algorithm is far more precise, it does not work well with sparse data, and delivers extremely low recall. On the other hand, the frequency algorithm provides excellent recall, but the results are not to be trusted since the precision is so low. All algorithms involving PMI require a much more sophisticated data structure, which necessitates excessively long processing and considerably more storage. Though it is less precise, the frequency algorithm is much faster and simpler. We show that overall, none of these algorithms performs well.

Expansion of WordNet Relationships. We extend the types of substitution words to include antonyms, meronyms of holonyms, and hyponyms of hypernyms, of the word to be substituted. We find that using the Fazly data set, there are situations where the hypernym $\rightarrow$ hyponym relationship improves precision, since it increases the set of substitutable words which, if the word-pair is compositional, are sometimes attested in the corpus, thereby reducing the number of mis-classified idioms. However, this does not appear to carry through to the second data set, which is not constrained to light verbs plus predicate nouns. We show that augmented substitutable word sets seem to improve precision, but do so at the cost of recall.

Substitutability as a criterion for identifying idioms. Our research is entirely predicated on the premise that substitutability is a suitable criterion for the identification of idioms. When alternative words can be substituted in a word-pair and found in the corpus, we consider the word-pair to be compositional and non-idiomatic. Every test performed in this study uses substitution of alternative words to discover non-compositional idioms.

However, the empirical evidence provided in this study shows that this assumption is wrong in two ways: failure to find substituted word-pairs in the corpus does not necessarily imply non-compositional idiomaticity; and successful discovery of substituted word-pairs does not mean that the word-pair is not an idiom. Our study shows several cases of word-pairs that are incorrectly classified as idioms simply because pairs created with substituted similar words do not occur in the corpus. Upon further examination, we observe that these word-pairs are simply tight collocations, not idioms. We also see idiomatic word-pairs for which substituted word-pairs are found in the corpus. This may be due to the fact that some idioms occur with slight variations (for example, *blow mind* and *blow head*), and because sometimes the words have an alternative sense which is compositional and can be substituted (such as *met match* and *met equal*, *lose touch* and *lose contact*, or *beaten track* and *beaten path*).

While substitutability may help to identify some tight collocations and very rigid non-compositional idioms, it is not an adequate criterion for identifying non-compositional idioms. Prior to this study, most of the research conducted relied on non-compositionality and substitutability to identify idioms. The work of Fazly (2007), a clear exception to this, shows the importance of applying lexical knowledge of idioms to the process of their identification. Nunberg et al. (1994) are correct in their suggestion that non-compositionality does not capture the essence of idiomaticity. This research clearly demonstrates that it is not a sufficient or necessary criterion.

6.2 Suggested future work

Expand test data. The Fazly data, used in these tests, is constrained to light verbs and nouns. The second data set is a small random extraction of word-pairs from Cowie et al. (1983). A more extensive set of word-pairs could be created by taking all word-pairs made up of nouns, adjectives, adverbs, and verbs within a distance of five words from the complete set of idioms presented by Cowie et al..

Expand data model. The data model is built using the BNC as a language sample. It would be interesting to use Google's Web 1T 5-gram data set (Brants and Franz, 2006) to build a language model. The words in this data set do not have POS tags, but a simplistic tagging algorithm, such as the one used by Justeson and Katz (1995) could be applied. The data is too sparse for some of our algorithms to work effectively. It would be interesting to discover whether the Google data set mitigates some of these problems. Alternatively, we could consider using a corpus of blogs which tend to be far more casual, such as the Blog Authorship Corpus (Schler et al., 2006), to build our model.

Switch from WordNet to a list of similar words. Throughout this experiment, we have used WordNet, which can be too broad or too narrow for our substitutional requirements. It would be interesting to use a *list of similar words* such as the one created by Lin (1998a) and used by Fazly (2007).

Expand classification criteria. Like Fazly (2007), it would be interesting to investigate and apply alternative linguistic cues to identify idiomaticity. The problem of determining those factors which can be combined with statistical measures to effectively identify idioms remains one of the challenges facing Computational Linguistics.

Appendix A

Input data

A.1 Stop words and BNC tags

Table A.1: Words that were excluded from the triples used in this experiment.

have	has	had	was
is	are	were	do
did	done	does	be
being	been	say	said
says	sais	doing	having
saying	must	may	shall
should	would	will	wo
sha	get	gets	also

Table A.2: Tags as described in the BNC documentation, and the new tags that are assigned to them for corpus processing. Only nouns, verbs, adjectives, and adverbs are included. All *being* and *having* verbs are ignored since they do not add semantic information.

Tag	*Description*	*New Tag*	*Example*
AJ0	Adjective (general or positive)	J	good, old, beautiful
AJC	Comparative adjective	J	better, older
AJS	Superlative adjective	J	best, oldest
AV0	General adverb: an adverb not subclassified as AVP or AVQ	R	often, well, longer, furthest.
AVP	Adverb particle	R	up, off, out
AVQ	Wh-adverb	R	when, where, how, why
NN0	Common noun, neutral for number	N	aircraft, data, committee
NN1	Singular common noun	N	pencil, goose, time
NN2	Plural common noun	N	pencils, geese, times
VVB	The finite base form of lexical verbs [Including the imperative and present subjunctive]	V	forget, send, live
VVD	The past tense form of lexical verbs	V	forgot, sent, lived
VVG	The -ing form of lexical verbs	V	forgetting, sending, living
VVI	The infinitive form of lexical verbs	V	forget, send, live
VVN	The past participle form of lexical verbs	V	forgotten, sent, lived
VVZ	The -s form of lexical verbs	V	forgets, sends, lives

A.2 Lists of word-pairs used in research

A.2.1 Development word-pairs

Table A.3: The Fazly training data set — a list of verb-noun word-pairs, including their frequency in the corpus and classification.

			Occurrence counts at distance						
Word-1	*POS*	*Word-2*	*tot*	*1*	*2*	*3*	*4*	*5*	*Idiom*
blow	VN	candle	7	0	4	2	0	1	
blow	VN	gaff	4	0	4	0	0	0	√
blow	VN	head	15	0	8	6	0	1	√
blow	VN	horn	16	0	15	1	0	0	√
blow	VN	smoke	11	3	2	3	3	0	√
blow	VN	top	12	0	10	2	0	0	√
blow	VN	whistle	24	1	23	0	0	0	√

continued on next page

Table A.3: *Fazly training data set (cont'd)*

			Occurrence counts at distance						
Word-1	*POS*	*Word-2*	*tot*	*1*	*2*	*3*	*4*	*5*	*Idiom*
bring	VN	bacon	12	0	0	11	0	1	√
bring	VN	bottle	36	1	23	6	6	0	
bring	VN	car	26	2	15	4	3	2	
bring	VN	flower	1	1	0	0	0	0	
bring	VN	luck	15	3	7	2	2	1	√
catch	VN	arm	3	0	1	0	2	0	
catch	VN	ball	17	0	12	4	1	0	
catch	VN	bug	1	0	0	1	0	0	√
catch	VN	cold	17	0	13	1	3	0	√
catch	VN	death	18	0	18	0	0	0	√
catch	VN	sight	56	50	5	1	0	0	√
cut	VN	branch	2	0	0	0	2	0	
cut	VN	bread	28	6	11	5	5	1	
cut	VN	corner	10	0	8	0	1	1	√
cut	VN	figure	28	0	3	18	5	2	√
cut	VN	loss	9	0	2	4	3	0	√
cut	VN	mustard	12	0	8	4	0	0	√
find	VN	bearing	1	0	1	0	0	0	√
find	VN	book	88	0	42	19	10	17	
find	VN	foot	6	0	0	1	2	3	√
find	VN	map	16	0	3	3	5	5	
find	VN	paper	35	2	8	9	6	10	
give	VN	bowl	6	0	0	5	1	0	
give	VN	ground	46	20	13	2	8	3	√
give	VN	medicine	25	2	5	10	3	5	
give	VN	money	453	49	110	211	54	29	
hit	VN	bottle	11	0	8	2	0	1	√
hit	VN	car	96	1	20	51	16	8	
hit	VN	fan	8	0	7	0	1	0	√
hit	VN	road	45	0	36	3	1	5	√
hit	VN	roof	25	1	21	2	1	0	√
hold	VN	arm	23	0	9	5	5	4	
hold	VN	bottle	1	0	0	0	1	0	
hold	VN	card	10	0	2	3	4	1	
hold	VN	dear	11	6	2	1	2	0	√
hold	VN	fort	15	0	15	0	0	0	√
hold	VN	knife	8	1	6	0	0	1	
hold	VN	pen	7	0	7	0	0	0	
hold	VN	ring	4	0	4	0	0	0	
hold	VN	sway	25	17	5	2	1	0	√

continued on next page

Table A.3: *Fazly training data set (cont'd)*

			Occurrence counts at distance						
Word-1	*POS*	*Word-2*	*tot*	*1*	*2*	*3*	*4*	*5*	*Idiom*
keep	VN	bird	5	0	2	2	1	0	
keep	VN	boat	20	0	11	2	3	4	
keep	VN	book	26	0	5	6	8	7	
keep	VN	fish	45	8	21	8	2	6	
keep	VN	journal	11	0	7	1	3	0	√
keep	VN	mind	272	1	161	66	27	17	√
keep	VN	pace	172	139	7	10	10	6	√
keep	VN	paper	24	1	8	3	6	6	
keep	VN	promise	48	0	42	6	0	0	√
keep	VN	spirit	13	1	4	4	2	2	√
keep	VN	temper	17	0	13	2	0	2	√
kick	VN	bucket	6	0	5	1	0	0	√
kick	VN	habit	32	0	27	5	0	0	√
kick	VN	heel	1	0	0	0	1	0	√
lay	VN	flower	1	0	0	1	0	0	
lay	VN	head	33	1	19	6	3	4	
lose	VN	cool	2	0	2	0	0	0	√
lose	VN	heart	26	21	4	1	0	0	√
lose	VN	land	7	0	3	1	1	2	
lose	VN	thread	4	0	4	0	0	0	√
make	VN	aeroplane	9	2	2	3	2	0	
make	VN	coffee	162	37	52	22	29	22	
make	VN	face	75	1	34	18	16	6	√
make	VN	fortune	87	0	65	16	4	2	√
make	VN	hay	22	14	4	3	1	0	√
make	VN	pastry	20	8	7	1	3	1	
make	VN	sandwich	23	0	5	12	4	2	
make	VN	scene	30	0	19	9	0	2	√
make	VN	tube	8	1	1	2	0	4	
move	VN	carriage	16	0	4	12	0	0	
place	VN	bag	18	0	1	5	10	2	
place	VN	bowl	45	1	3	10	18	13	
pull	VN	arm	7	0	4	3	0	0	
pull	VN	plug	30	0	27	3	0	0	√
pull	VN	punch	2	0	0	0	0	2	√
pull	VN	string	10	0	6	4	0	0	√
push	VN	button	36	7	14	8	3	4	
push	VN	chair	6	0	3	2	1	0	
push	VN	plate	3	0	2	0	1	0	
put	VN	book	95	1	34	26	23	11	

continued on next page

Table A.3: *Fazly training data set (cont'd)*

Word-1	*POS*	*Word-2*	*Occurrence counts at distance* tot	*1*	*2*	*3*	*4*	*5*	*Idiom*
put	VN	bottle	52	0	21	11	9	11	
put	VN	finger	229	1	204	13	7	4	√
put	VN	jacket	43	0	11	8	10	14	
put	VN	package	46	0	5	20	15	6	
see	VN	baby	58	2	32	14	7	3	
see	VN	star	18	2	8	2	2	4	√
set	VN	cup	27	0	10	6	4	7	
set	VN	foot	172	139	10	12	8	3	√
set	VN	pace	70	0	39	17	4	10	√
set	VN	sail	10	9	0	1	0	0	√
take	VN	bottle	32	0	15	6	7	4	
take	VN	hammer	4	0	4	0	0	0	
take	VN	jacket	26	0	15	10	0	1	
take	VN	rap	12	0	11	0	1	0	√
take	VN	root	50	40	1	6	3	0	√
throw	VN	book	13	0	7	5	1	0	
throw	VN	egg	1	0	1	0	0	0	
touch	VN	base	3	1	1	1	0	0	√
touch	VN	cheek	14	0	12	0	1	1	
touch	VN	nerve	3	0	1	2	0	0	√
touch	VN	wood	14	13	1	0	0	0	√

A.2.2 Test data sets

Table A.4: The Fazly test data set – a list of verb-noun word-pairs, including their frequency in the corpus and classification.

Word-1	*POS*	*Word-2*	*Occurrence counts at distance* tot	*1*	*2*	*3*	*4*	*5*	*Idiom*
blow	VN	bridge	7	0	1	4	2	0	
blow	VN	fuse	3	0	3	0	0	0	√
blow	VN	gasket	2	0	1	0	1	0	√
blow	VN	hole	6	0	4	1	1	0	√
blow	VN	mind	4	0	3	0	1	0	√
blow	VN	trumpet	19	0	5	13	0	1	√
bring	VN	bag	17	0	9	7	1	0	
bring	VN	cup	29	1	5	14	8	1	
catch	VN	attention	28	1	20	3	3	1	√
catch	VN	breath	79	1	75	1	2	0	√

continued on next page

Table A.4: *Fazly test data set (cont'd)*

			Occurrence counts at distance						
Word-1	*POS*	*Word-2*	*tot*	*1*	*2*	*3*	*4*	*5*	*Idiom*
catch	VN	fire	43	39	3	0	1	0	√
catch	VN	horse	6	0	4	1	0	1	
catch	VN	imagination	6	0	4	2	0	0	√
catch	VN	rabbit	4	0	2	1	0	1	
catch	VN	trout	9	2	5	2	0	0	
cut	VN	cake	47	0	18	14	7	8	
cut	VN	cloth	27	2	9	4	8	4	√
cut	VN	cord	15	0	5	4	6	0	√
cut	VN	dash	13	0	9	3	1	0	√
cut	VN	grass	53	13	31	7	1	1	
cut	VN	hand	32	1	14	10	2	5	
cut	VN	rate	51	0	5	29	13	4	√
cut	VN	rope	16	0	10	3	2	1	
cut	VN	throat	42	2	31	8	1	0	√
cut	VN	tree	22	0	4	9	7	2	
cut	VN	wire	14	0	2	5	6	1	
cut	VN	wood	35	8	4	7	11	5	
find	VN	bottle	12	0	11	1	0	0	
find	VN	box	22	0	9	8	1	4	
find	VN	tongue	4	0	3	1	0	0	√
give	VN	birth	130	126	1	1	1	1	√
give	VN	drink	34	1	6	20	5	2	
give	VN	drug	14	0	1	6	5	2	
give	VN	flick	3	0	0	2	1	0	√
give	VN	gift	28	0	8	9	6	5	
give	VN	land	31	7	6	5	8	5	
give	VN	lift	160	1	9	129	11	10	√
give	VN	mug	4	0	1	2	1	0	
give	VN	notice	248	112	34	47	36	19	√
give	VN	push	29	0	0	12	13	4	√
give	VN	sack	12	0	0	10	0	2	√
give	VN	slip	14	0	0	8	4	2	√
give	VN	ticket	17	0	2	9	5	1	
give	VN	way	602	548	2	16	17	19	√
give	VN	whirl	6	0	0	5	1	0	√
hit	VN	ceiling	10	0	7	2	0	1	√
hit	VN	deck	14	0	14	0	0	0	√
hit	VN	headline	1	0	1	0	0	0	√
hit	VN	jackpot	28	1	27	0	0	0	√
hit	VN	man	48	20	16	8	2	2	

continued on next page

Table A.4: *Fazly test data set (cont'd)*

			Occurrence counts at distance						
Word-1	*POS*	*Word-2*	*tot*	*1*	*2*	*3*	*4*	*5*	*Idiom*
hit	VN	spot	17	1	5	6	1	4	√
hit	VN	wall	55	0	37	17	1	0	√
hold	VN	baby	18	0	13	3	1	1	
hold	VN	bird	4	0	4	0	0	0	
hold	VN	bowl	1	0	1	0	0	0	
hold	VN	fire	21	5	10	3	2	1	√
hold	VN	ground	19	0	12	3	2	2	√
hold	VN	hand	168	0	110	31	10	17	√
hold	VN	horse	7	0	4	1	1	1	√
hold	VN	key	64	6	50	4	4	0	
hold	VN	plate	3	0	1	1	1	0	
hold	VN	tongue	29	0	26	1	0	2	√
hold	VN	tray	1	0	0	1	0	0	
keep	VN	cool	13	0	9	4	0	0	√
keep	VN	end	30	0	11	3	6	10	√
keep	VN	grip	22	0	10	11	1	0	√
keep	VN	hand	61	2	30	12	9	8	√
keep	VN	head	138	1	90	25	14	8	√
keep	VN	horse	26	0	21	1	2	2	
keep	VN	pig	4	0	3	1	0	0	
keep	VN	secret	133	0	44	56	23	10	√
keep	VN	tab	1	0	0	1	0	0	√
keep	VN	watch	94	35	26	27	5	1	√
keep	VN	word	44	0	36	5	2	1	√
lay	VN	block	1	0	0	0	1	0	
lay	VN	carpet	12	1	1	4	5	1	
lay	VN	pipe	6	1	1	0	3	1	
lay	VN	waste	9	8	1	0	0	0	√
lose	VN	deposit	2	0	2	0	0	0	
lose	VN	face	28	24	1	2	0	1	√
lose	VN	ground	9	7	1	0	0	1	√
lose	VN	head	22	1	18	0	0	3	√
lose	VN	home	25	0	20	1	1	3	
lose	VN	money	114	68	18	13	11	4	
lose	VN	rag	6	0	6	0	0	0	√
lose	VN	shirt	1	0	1	0	0	0	√
lose	VN	temper	83	0	80	3	0	0	√
lose	VN	touch	54	46	3	2	1	2	√
make	VN	beeline	4	0	4	0	0	0	√
make	VN	biscuit	2	0	1	0	1	0	

continued on next page

Table A.4: *Fazly test data set (cont'd)*

			Occurrence counts at distance						
Word-1	*POS*	*Word-2*	*tot*	*1*	*2*	*3*	*4*	*5*	*Idiom*
make	VN	cake	47	2	18	16	2	9	
make	VN	custard	15	4	6	3	1	1	
make	VN	debut	142	2	70	54	10	6	√
make	VN	history	72	23	18	17	8	6	√
make	VN	hit	10	1	4	3	0	2	√
make	VN	killing	29	0	23	5	0	1	√
make	VN	mark	113	1	89	17	4	2	√
make	VN	peace	111	64	29	8	3	7	√
make	VN	pie	17	0	4	10	2	1	
make	VN	pile	10	0	9	0	0	1	√
make	VN	plastic	9	0	2	4	2	1	
make	VN	scone	1	0	1	0	0	0	
make	VN	toy	6	0	1	3	1	1	
move	VN	car	33	0	23	5	3	2	
move	VN	house	121	51	4	23	26	17	√
move	VN	mountain	3	0	3	0	0	0	√
pull	VN	box	2	0	0	0	1	1	
pull	VN	chain	7	0	7	0	0	0	√
pull	VN	chair	11	0	4	7	0	0	
pull	VN	finger	7	0	5	1	0	1	√
pull	VN	hair	22	1	13	3	4	1	√
pull	VN	leg	33	0	8	21	3	1	√
pull	VN	shirt	9	0	4	1	3	1	
pull	VN	weight	23	0	18	2	1	2	√
push	VN	barrow	2	0	0	1	0	1	
push	VN	bike	6	0	5	0	1	0	
push	VN	boat	5	0	3	1	1	0	√
push	VN	luck	30	0	30	0	0	0	√
push	VN	paper	3	1	1	0	1	0	√
push	VN	trolley	4	0	1	1	1	1	
put	VN	box	132	2	15	19	51	45	
put	VN	candle	15	0	7	8	0	0	
put	VN	car	127	0	43	21	39	24	
put	VN	flesh	20	8	10	1	1	0	√
put	VN	gloss	11	0	6	3	0	2	√
put	VN	helmet	14	0	6	6	0	2	
put	VN	key	42	0	35	4	3	0	
see	VN	daylight	16	8	3	2	3	0	√
see	VN	red	16	4	3	5	1	3	√
see	VN	sight	18	0	1	6	7	4	√

continued on next page

Table A.4: *Fazly test data set (cont'd)*

Word-1	*POS*	*Word-2*	*Occurrence counts at distance* *tot*	*1*	*2*	*3*	*4*	*5*	*Idiom*
see	VN	woman	105	1	46	21	20	17	
set	VN	cap	5	0	3	1	0	1	√
set	VN	carriage	11	1	4	5	1	0	
set	VN	fire	335	180	76	34	35	10	√
set	VN	stage	49	0	21	9	8	11	√
set	VN	tank	32	0	1	9	14	8	
shoot	VN	bolt	2	0	1	1	0	0	√
smell	VN	rat	13	0	13	0	0	0	√
take	VN	air	66	7	10	32	7	10	√
take	VN	arm	37	1	18	7	6	5	
take	VN	biscuit	8	0	6	1	0	1	√
take	VN	boat	54	0	30	7	11	6	
take	VN	box	39	2	12	8	10	7	
take	VN	ease	6	0	5	0	1	0	√
take	VN	folder	1	0	1	0	0	0	
take	VN	gun	21	0	12	6	0	3	
take	VN	handkerchief	6	0	4	1	1	0	
take	VN	heart	102	46	13	12	16	15	√
take	VN	lunch	71	17	11	16	18	9	
take	VN	notebook	6	0	2	3	0	1	
take	VN	plate	12	0	7	2	2	1	
take	VN	prize	27	0	10	9	6	2	
throw	VN	brick	3	0	3	0	0	0	
throw	VN	hat	3	0	2	0	0	1	
throw	VN	towel	23	0	2	21	0	0	
touch	VN	finger	1	0	0	0	0	1	
touch	VN	forehead	2	0	1	0	1	0	
touch	VN	shoulder	9	0	4	1	2	2	

Table A.5: The Cowie test data set – a list of word-pairs not constrained to verb-noun pairs, including their frequency in the corpus and classification.

Word-1	*POS*	*Word-2*	*Occurrence counts at distance* *tot*	*1*	*2*	*3*	*4*	*5*	*Idiom*
cut	VR	loose	7	5	2	0	0	0	√
cut	VV	make	43	0	3	10	13	17	
cut	VV	pasted	6	0	3	2	1	0	√
cut	VV	use	11	0	0	1	6	4	
darken	VN	colour	4	0	3	0	0	1	
darken	VN	door	4	0	3	0	0	1	√

continued on next page

Table A.5: *Cowie data set (cont'd)*

			Occurrence counts at distance						
Word-1	*POS*	*Word-2*	*tot*	*1*	*2*	*3*	*4*	*5*	*Idiom*
diamond	NN	cluster	21	11	8	2	0	0	
diamond	NN	water	1	0	0	1	0	0	√
dog	NN	day	23	3	1	8	9	2	√
dog	NN	fleas	77	72	0	2	1	2	
dog	NN	leg	12	3	0	3	2	4	
dog	NN	life	28	18	1	4	3	2	√
double	JN	bonds	12	12	0	0	0	0	
double	RV	glazed	26	26	0	0	0	0	
double	RV	take	16	15	0	1	0	0	√
double	JN	talk	7	7	0	0	0	0	√
drink	VN	beer	38	14	7	7	7	3	
drink	VN	fish	3	0	1	2	0	0	√
drive	VN	bargain	15	0	0	12	2	1	√
drive	VN	vehicles	11	1	5	3	2	0	
drop	NN	bucket	4	0	0	4	0	0	√
drop	NN	level	14	0	0	7	4	3	
drown	VN	man	1	0	0	1	0	0	
drown	VN	sorrow	1	0	0	1	0	0	√
dry	JN	dust	15	3	6	2	3	1	√
dry	JN	ground	22	16	3	0	1	2	
eat	VN	dog	18	14	1	0	3	0	√
eat	VN	steak	11	2	4	3	1	1	
end	NN	episode	19	0	2	11	5	1	
end	NN	road	196	12	7	151	21	5	√
explore	VN	avenue	4	0	3	0	1	0	√
explore	VN	detail	14	0	4	4	3	3	
far	JN	cry	177	176	1	0	0	0	√
far	JN	shore	13	11	2	0	0	0	
feather	NN	cap	18	0	1	13	3	1	√
feather	NN	mattress	12	12	0	0	0	0	
flat	JN	board	10	4	0	3	1	2	√
flat	JN	lands	13	9	3	1	0	0	
flight	NN	fancy	18	0	16	2	0	0	√
flight	NN	instruments	16	16	0	0	0	0	
flying	JN	bird	37	34	1	0	1	1	
flying	JN	colours	44	44	0	0	0	0	√
fresh	JN	daisy	6	0	0	5	1	0	√
fresh	JN	ingredients	13	11	1	1	0	0	
funny	JN	business	21	18	2	0	1	0	√
funny	JN	joke	31	29	1	1	0	0	

continued on next page

Table A.5: *Cowie data set (cont'd)*

			Occurrence counts at distance						
Word-1	*POS*	*Word-2*	*tot*	*1*	*2*	*3*	*4*	*5*	*Idiom*
gentleman	NN	agreement	9	8	0	0	0	1	√
gentleman	NN	position	16	2	1	0	5	8	
give	VN	gift	16	0	0	8	2	6	
give	VN	inch	18	0	9	6	2	1	√
golden	JN	hair	43	42	0	0	1	0	
golden	JN	opportunity	78	76	1	1	0	0	√
heart	NN	heart	21	0	7	4	7	3	√
heart	NN	lungs	47	0	40	6	1	0	
hit	VN	bottle	11	0	8	2	0	1	√
hit	VN	woman	89	2	80	0	2	5	
jet	NN	plane	14	12	0	1	1	0	
jet	NN	set	11	11	0	0	0	0	√
kill	VN	bacteria	19	5	8	3	2	1	
kill	VN	time	29	11	5	8	2	3	√
kiss	NN	death	24	0	22	1	1	0	√
kiss	NN	lips	24	0	0	11	10	3	
know	VN	places	47	2	7	14	14	10	
know	VN	ropes	8	1	6	1	0	0	√
lame	JN	duck	17	17	0	0	0	0	√
lame	JN	leg	17	17	0	0	0	0	
met	VN	match	20	0	16	1	1	2	√
met	VN	mother	34	0	21	4	6	3	
mine	NN	coal	11	4	0	1	3	3	
mine	NN	information	18	1	13	3	0	1	√
old	JN	hills	19	0	2	15	0	2	√
old	JN	places	21	9	1	4	4	3	
pig	NN	ear	11	11	0	0	0	0	√
pig	NN	farmer	12	9	0	1	1	1	
play	VN	instruments	23	3	11	2	4	3	
play	VN	possum	1	1	0	0	0	0	√
pound	VN	beat	2	1	0	0	1	0	√
pound	VN	door	2	1	0	0	1	0	
rags	NN	dirt	22	0	19	2	0	1	
rags	NN	riches	22	0	19	2	0	1	√
rain	VN	cats	1	1	0	0	0	0	√
rain	VN	umbrella	1	1	0	0	0	0	
rat	NN	race	27	27	0	0	0	0	√
rat	NN	stomach	17	16	0	0	0	1	
red	JN	brick	140	132	5	1	0	2	
red	JN	bus	11	5	2	2	1	1	

continued on next page

Table A.5: *Cowie data set (cont'd)*

			Occurrence counts at distance						
Word-1	*POS*	*Word-2*	*tot*	*1*	*2*	*3*	*4*	*5*	*Idiom*
red	JN	carpet	48	43	1	1	3	0	√
red	JN	herring	56	56	0	0	0	0	√
red	JN	houses	11	3	3	3	0	2	
red	JN	tape	162	162	0	0	0	0	√
short	JJ	straight	17	11	2	2	1	1	
short	JJ	sweet	7	0	5	0	2	0	√
shot	NN	dark	11	0	0	11	0	0	√
shot	NN	target	11	0	7	3	0	1	
sit	VR	comfortably	40	29	7	3	1	0	
sit	VR	tight	8	8	0	0	0	0	√
sob	NN	story	9	9	0	0	0	0	√
sob	NN	throat	13	0	0	2	10	1	
son	NN	bitch	30	0	0	30	0	0	√
son	NN	gun	6	0	0	6	0	0	√
son	NN	mother	30	0	5	11	11	3	
son	NN	years	28	0	4	7	7	10	
spit	NN	ground	6	0	6	0	0	0	
spit	NN	polish	6	0	6	0	0	0	√
straight	JN	answer	29	27	0	2	0	0	√
straight	JN	line	13	11	2	0	0	0	
stuffed	JN	bird	14	12	2	0	0	0	
stuffed	JN	shirt	4	4	0	0	0	0	√
sweat	NN	blood	10	1	6	1	1	1	√
sweat	NN	face	38	0	1	16	11	10	
swing	VN	arm	50	0	34	15	1	0	
swing	VN	cat	5	0	5	0	0	0	√
take	VN	houses	13	1	3	7	2	0	
take	VN	powder	5	0	4	0	1	0	√
tall	JN	tale	3	3	0	0	0	0	√
tall	JN	tower	27	10	13	4	0	0	
tempt	VN	fate	10	10	0	0	0	0	√
tempt	VN	person	10	10	0	0	0	0	
think	NN	idea	33	33	0	0	0	0	
think	NN	tank	33	33	0	0	0	0	√
time	NV	convince	11	1	4	3	2	1	
time	NV	tell	203	1	137	20	23	22	√
touch	VN	ground	24	0	24	0	0	0	
touch	VN	wood	14	13	1	0	0	0	√
true	JJ	accurate	13	0	13	0	0	0	
true	JJ	blue	17	17	0	0	0	0	√

continued on next page

Table A.5: *Cowie data set (cont'd)*

Word-1	*POS*	*Word-2*	*Occurrence counts at distance* tot	*1*	*2*	*3*	*4*	*5*	*Idiom*
twinkling	NN	eye	16	0	0	15	1	0	√
twinkling	NN	stars	16	0	0	15	1	0	
ugly	JN	face	32	26	2	1	3	0	
ugly	JN	sin	5	0	5	0	0	0	√
warm	JN	climate	18	10	5	2	1	0	
warm	JN	toast	10	4	6	0	0	0	√
watch	VN	clock	12	0	4	6	1	1	√
watch	VN	films	21	6	11	1	3	0	
wet	JN	blanket	18	17	0	1	0	0	√
wet	JN	road	12	9	0	2	1	0	
wheeling	VV	dealing	15	0	15	0	0	0	√
wheeling	VV	stopping	15	0	15	0	0	0	
whipping	JN	boy	8	8	0	0	0	0	√
whipping	JN	slave	8	8	0	0	0	0	
white	JN	elephant	37	36	1	0	0	0	√
white	JN	houses	22	12	8	1	0	1	
white	JN	lie	15	15	0	0	0	0	√
white	JN	sand	22	12	8	1	0	1	
wild	JN	dog	13	12	1	0	0	0	
wild	JN	oats	11	11	0	0	0	0	√
wind	NN	change	26	1	21	3	1	0	√
wind	NN	rain	22	0	12	7	1	2	

Appendix B

Research results

The following sections provide tables and charts detailing the results of this study. On many charts a black horizontal line represents the baseline: the PMI calculation using a bag of words substituting synonyms only.

B.1 All factors using synonyms only

Figures B.1 to B.5 illustrate the results.

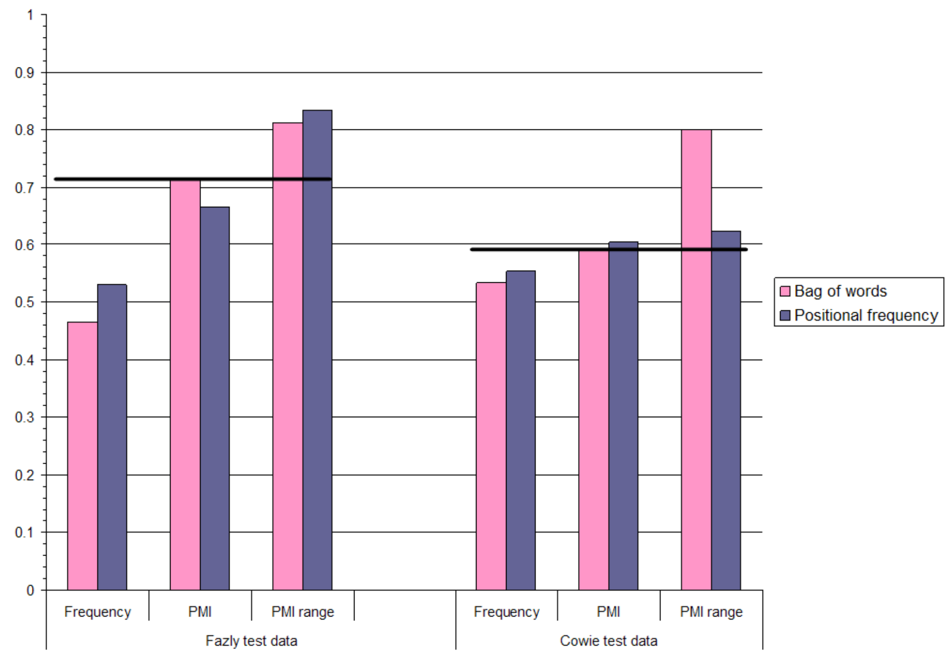

Figure B.1: The precision measured across both data sets using all three algorithms.

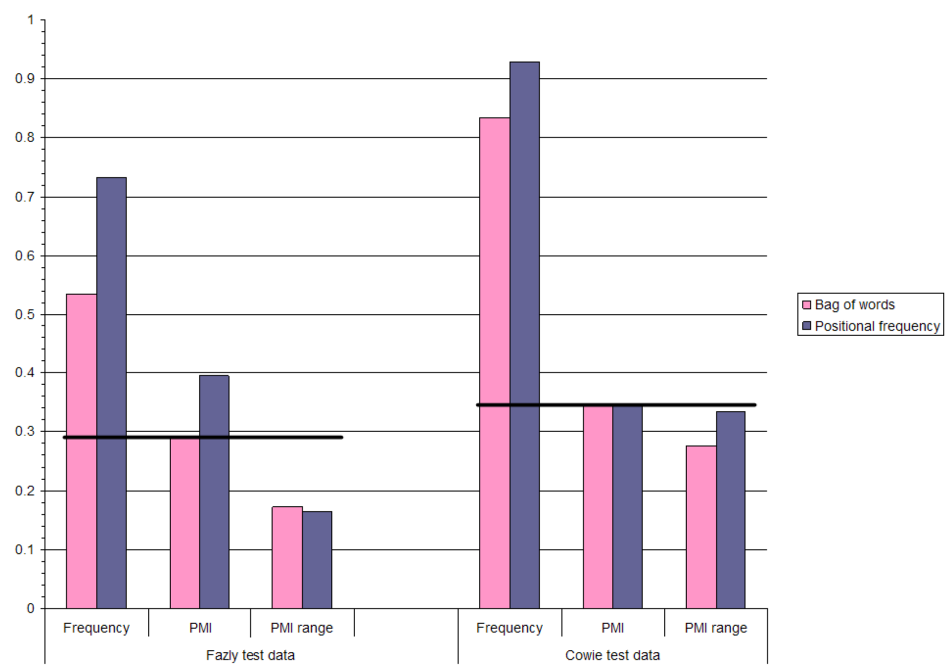

Figure B.2: The recall measured across both data sets using all three algorithms.

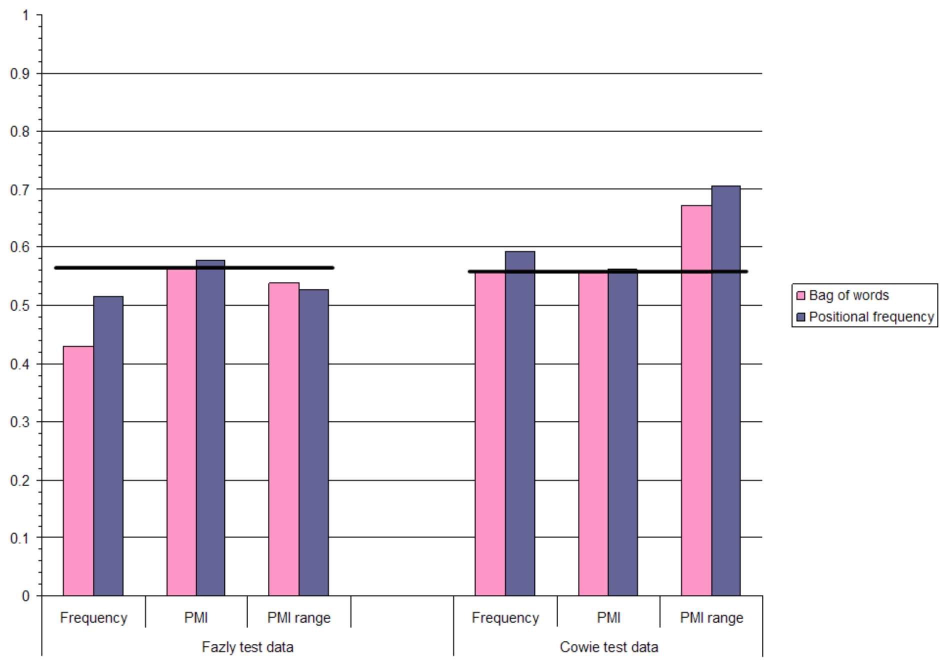

Figure B.3: The accuracy measured across both data sets using all three algorithms.

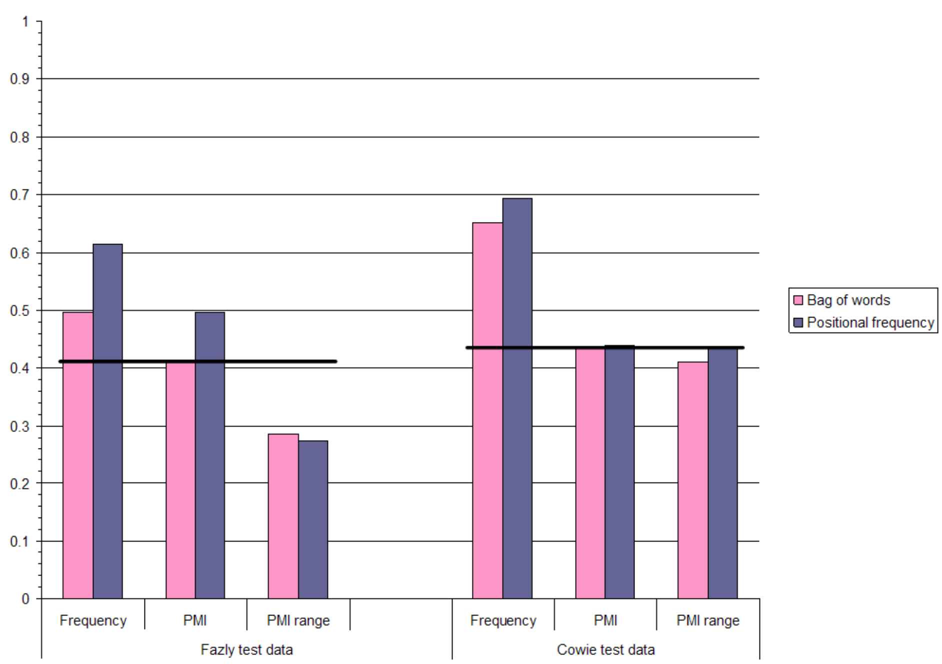

Figure B.4: The F-score measured across both data sets using all three algorithms.

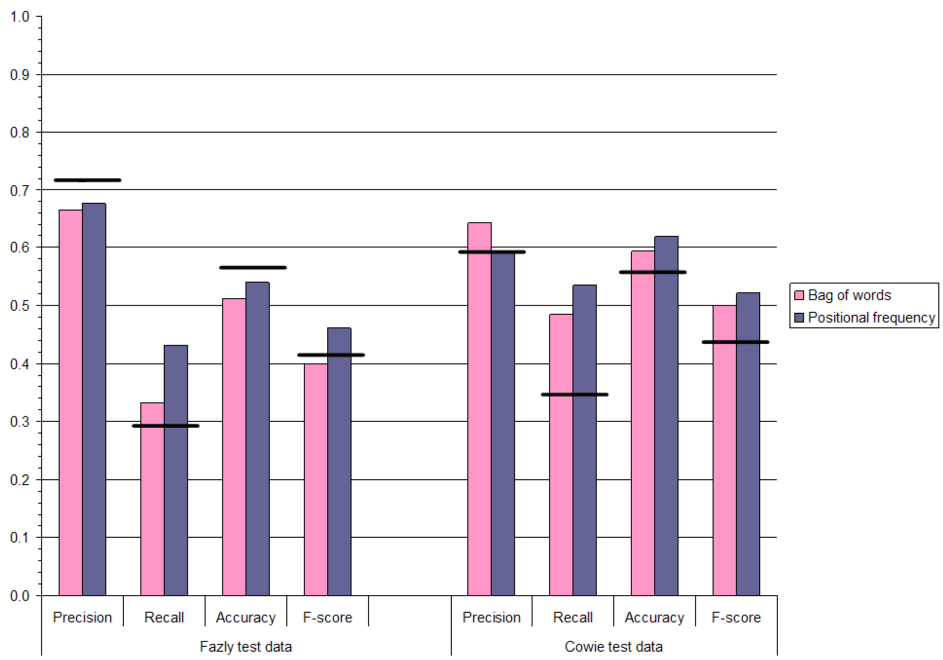

Figure B.5: All measures across both data sets using the average performance of the three algorithms.

B.2 WordNet relationships

Figures B.6 to B.10 show the effect of additional WordNet relationships on our tests involving both the Fazly test data and the Cowie test data. In each of these figures S = synonyms; A = antonyms; M = holonym → meronym; and H = hypernym → hyponym.

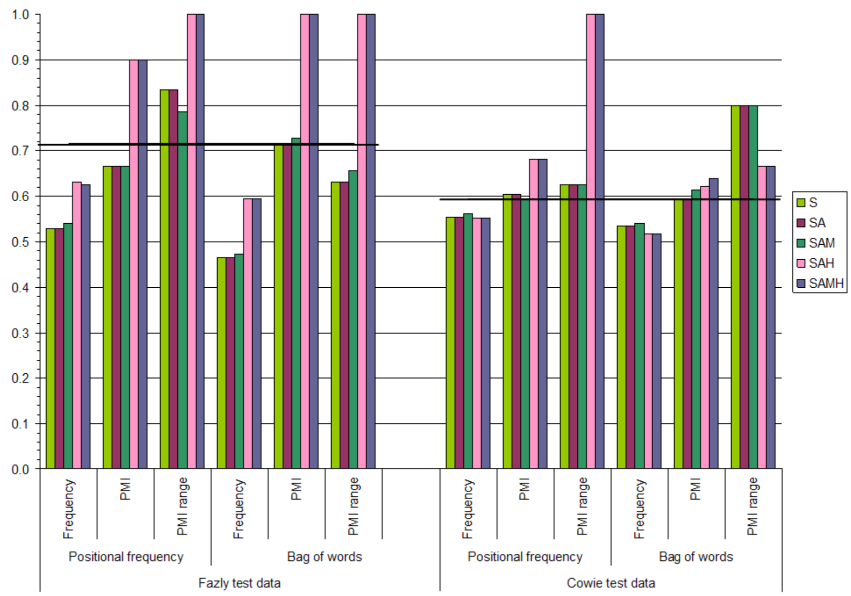

Figure B.6: The precision measured across both data sets.

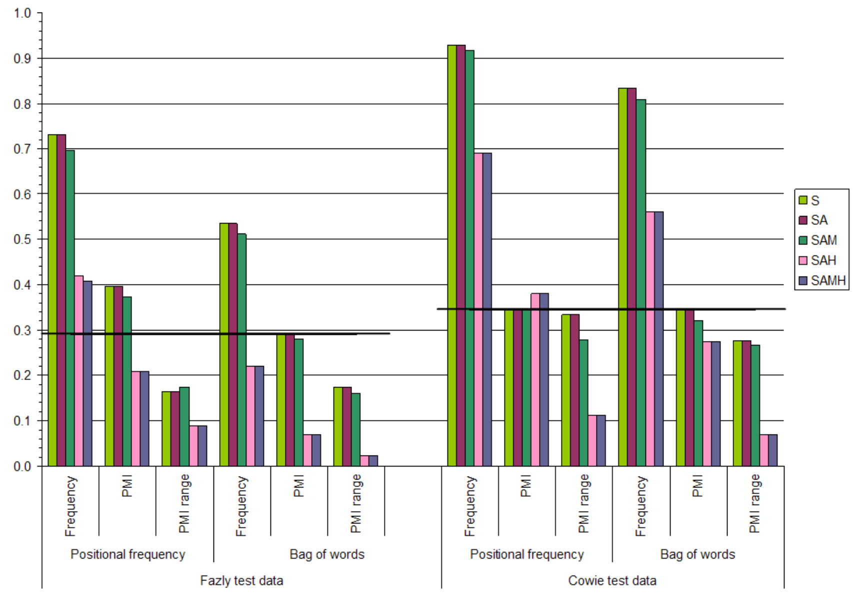

Figure B.7: The recall measured across both data sets.

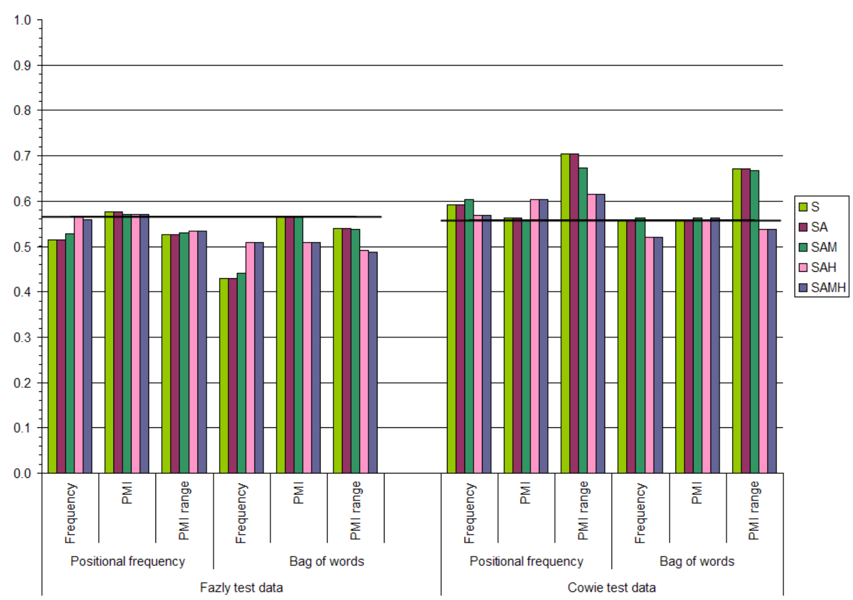

Figure B.8: The accuracy measured across both data sets.

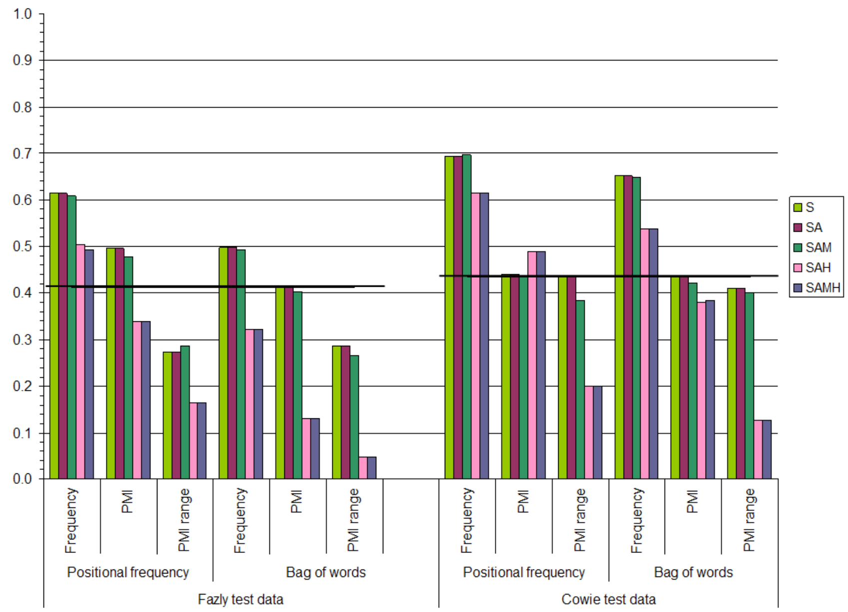

Figure B.9: The F-score measured across both data sets.

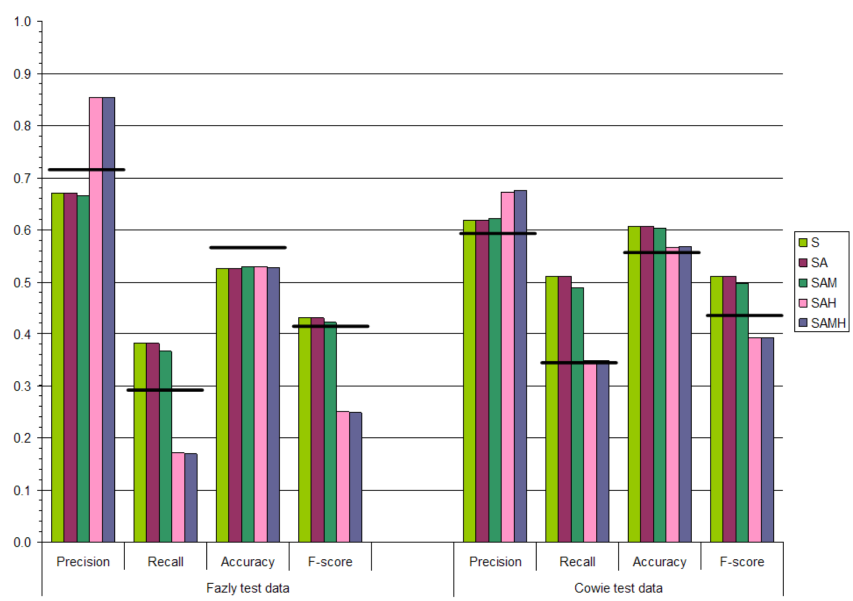

Figure B.10: All measures averaged across both data sets.

B.3 Raw results of evaluations

The following tables provide the evaluations which resulted from all algorithms when substituting synonyms only. A $\surd$ character indicates that the word-pair was classified as an idiom. A dash (-) indicates that it was not possible to classify the word-pair due to low occurrence frequency (as discussed in Section 3.3.3).

Table B.1: The raw results for the Fazly test data set for all algorithms substituting synonyms only.

		Classifications					
		Positional			*Bag-of-words*		
Word-1	*Word-2*	Frequency	*PMI*	*PMI range*	Frequency	*PMI*	*PMI range*
Non-idioms							
blow	bridge	$\surd$		-	$\surd$		-
bring	bag						
bring	cup	$\surd$					
catch	horse						
catch	insect	$\surd$	$\surd$	-	$\surd$		
catch	rabbit	$\surd$		-			
catch	trout	$\surd$	$\surd$	-	$\surd$	$\surd$	$\surd$
cut	cake	$\surd$	$\surd$	-	$\surd$		
cut	grass	$\surd$			$\surd$		
cut	hand	$\surd$			$\surd$		
cut	rope	$\surd$	$\surd$	-	$\surd$		
cut	tree	$\surd$			$\surd$		
cut	wire	$\surd$		-	$\surd$		-
cut	wood	$\surd$		-	$\surd$		
find	bottle						
find	box						
give	drink	$\surd$			$\surd$		
give	drug	$\surd$			$\surd$		
give	gift						
give	land						
give	mug						
give	ticket	$\surd$			$\surd$		
hit	man		$\surd$				
hold	baby						
hold	bird						
hold	bowl						
hold	key	$\surd$	$\surd$		$\surd$		
hold	plate						

continued on next page

Table B.1: *Raw results — Fazly test data set (cont'd)*

		Classifications					
		Positional			*Bag-of-words*		
Word-1	*Word-2*	Frequency	*PMI*	*PMI range*	Frequency	*PMI*	*PMI range*
hold	tray						
keep	horse	√			√		
keep	pig	√	√	-	√	√	
lay	block						
lay	carpet	√			√		
lay	pipe	√		-			
lose	deposit	√		-	√		-
lose	home	√	√		√		
lose	money	√		-	√	√	√
lose	ticket	√		-	√		-
make	biscuit						
make	cake	√			√		
make	custard	√		-	√		-
make	pancake	√		-	√		-
make	pie	√		-	√		
make	plastic	√			√		
make	scone	√	√	-	√		-
make	toy	√			√		
move	car						
pull	box	√			√		
pull	chair	√	√		√	√	
pull	shirt	√			√		
push	barrow	√		-	√	√	-
push	bike	√			√		
push	trolley	√		-	√	√	-
put	box	√			√		
put	candle	√	√		√		
put	car	√			√		
put	helmet	√		-	√	√	-
put	key	√	√		√		
see	woman	√			√		
set	carriage	√		-	√		
set	tank				√		
take	arm	√			√		
take	boat	√			√		
take	box				√		
take	folder	√					
take	gun						
take	handkerchief	√		-	√		

continued on next page

Table B.1: *Raw results — Fazly test data set (cont'd)*

		Classifications					
		Positional			*Bag-of-words*		
Word-1	*Word-2*	Frequency	*PMI*	*PMI range*	Frequency	*PMI*	*PMI range*
take	lunch	√			√		
take	notebook	√		-	√		
take	plate						
take	prize	√			√		
throw	brick	√					
throw	hat		√	√			
throw	towel	√	√	-	√		
touch	finger	√	√	-	√	√	
touch	forehead	√	√	-	√	√	-
touch	shoulder	√	√	√	√	√	√
Idioms							
blow	fuse	√		-	√		-
blow	gasket	√		-	√		-
blow	hole	√		-	√	√	-
blow	mind	√					
blow	trumpet	√	√	-	√	√	√
catch	attention	√			√		
catch	breath	√	√	√	√	√	√
catch	fancy	√	√	-	√	√	
catch	fire	√			√		
catch	imagination						
cut	cloth	√	√		√	√	√
cut	cord	√		-	√	√	-
cut	dash	√	√	-	√	√	√
cut	rate		√				
cut	throat	√			√		
find	tongue						
give	birth	√	√	-	√	√	√
give	creep	√	√		√	√	
give	flick						
give	lift	√	√				
give	notice	√	√	-	√	√	
give	push	√	√			√	√
give	sack						
give	slip						
give	way	√	√	√			
give	whirl						
hit	ceiling		√				

continued on next page

Table B.1: *Raw results — Fazly test data set (cont'd)*

		Classifications					
		Positional			*Bag-of-words*		
Word-1	*Word-2*	Frequency	*PMI*	*PMI range*	Frequency	*PMI*	*PMI range*
hit	deck	$\surd$		-	$\surd$		
hit	headline	$\surd$	$\surd$	$\surd$		$\surd$	$\surd$
hit	jackpot	$\surd$	$\surd$	-	$\surd$	$\surd$	-
hit	spot						
hit	wall	$\surd$					
hold	fire						
hold	ground						
hold	hand	$\surd$			$\surd$	$\surd$	$\surd$
hold	horse						
hold	tongue	$\surd$					
keep	cool	$\surd$		-	$\surd$		-
keep	end						
keep	grip	$\surd$					
keep	hand						
keep	head						
keep	secret	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$
keep	tab	$\surd$	$\surd$	$\surd$		$\surd$	$\surd$
keep	watch	$\surd$			$\surd$		
keep	word	$\surd$	$\surd$		$\surd$		
lay	waste	$\surd$			$\surd$		
lose	face	$\surd$	$\surd$	-			
lose	ground	$\surd$	$\surd$	$\surd$	$\surd$		
lose	head	$\surd$					
lose	rag	$\surd$	$\surd$	-	$\surd$	$\surd$	-
lose	shirt	$\surd$		-	$\surd$		-
lose	temper	$\surd$		-	$\surd$	$\surd$	$\surd$
lose	touch	$\surd$	$\surd$	$\surd$	$\surd$		
make	beeline	$\surd$		-	$\surd$	$\surd$	-
make	debut	$\surd$	$\surd$		$\surd$		
make	history	$\surd$			$\surd$		
make	hit				$\surd$		
make	killing	$\surd$			$\surd$		
make	mark	$\surd$			$\surd$		
make	peace	$\surd$			$\surd$		
make	pile						
move	house	$\surd$	$\surd$	$\surd$		$\surd$	
move	mountain	$\surd$	$\surd$	-			
pull	chain						
pull	finger	$\surd$			$\surd$		

continued on next page

Table B.1: *Raw results — Fazly test data set (cont'd)*

		Classifications					
		Positional			*Bag-of-words*		
Word-1	*Word-2*	Frequency	*PMI*	*PMI range*	Frequency	*PMI*	*PMI range*
pull	hair	√			√		
pull	leg	√	√		√	√	
pull	weight	√		-	√	√	√
push	boat	√	√	-	√		
push	luck	√	√	-	√	√	√
push	paper	√		-	√		
put	flesh	√	√				
put	gloss		√	√			
see	daylight	√	√	-			
see	red	√					
see	sight						
set	cap						
set	fire	√	√	√	√		
set	stage	√	√				
shoot	bolt	√		-	√	√	-
smell	rat	√	√	-	√	√	-
take	air						
take	biscuit	√			√		
take	ease						
take	heart	√					

Table B.2: The raw results for the Cowie test data set for all algorithms substituting synonyms only.

		Classifications					
		Positional			*Bag-of-words*		
Word-1	*Word-2*	Frequency	*PMI*	*PMI range*	Frequency	*PMI*	*PMI range*
Non-idioms							
above	effect	√		-	√	√	
above	terms						
alive	died	√	√	-	√	√	-
bag	feet	√	√	-	√		
bat	ball	√		-	√		-
beaten	egg	√		-	√		-
bird	fly	√			√		
bitter	chocolate	√	√	-	√	√	√
blaze	fire	√		-	√		-
cut	give						
cut	here	√		-	√		
cut	make	√			√		

continued on next page

Table B.2: *Raw results — Cowie test data set (cont'd)*

		Classifications					
		Positional			*Bag-of-words*		
Word-1	*Word-2*	Frequency	*PMI*	*PMI range*	Frequency	*PMI*	*PMI range*
cut	use	√					
darken	colour	√		-	√		-
diamond	cluster	√		-	√		-
dog	fleas	√		-	√		-
dog	leg	√	√	-	√		
double	bonds	√		-	√	√	-
double	glazed	√		-	√		-
drink	beer	√		-	√		-
drive	vehicles	√	√	-			
drop	level						
drown	man			-	√	√	-
dry	ground						
eat	steak	√	√	-	√	√	
end	episode						
explore	detail	√	√	√	√	√	
far	shore	√		-	√		-
feather	mattress	√		-	√		-
flat	lands	√		-	√		-
flight	instruments	√		-	√		-
flying	bird						
fresh	ingredients	√	√	-	√	√	-
funny	joke				√		
gentleman	position						
give	gift						
golden	hair	√		-	√		-
heart	lungs	√		-	√	√	-
hit	woman						
jet	plane	√		-	√		-
kill	bacteria	√		-	√		-
kiss	lips	√		-	√		-
know	places	√			√		
lame	leg			-	√		
met	mother	√		-	√		-
mine	coal	√		-	√		-
old	places	√		-	√		-
pig	farmer	√		-	√	√	-
play	instruments	√	√	-	√	√	
pound	door						
rat	stomach	√		-	√	√	-

continued on next page

Table B.2: *Raw results — Cowie test data set (cont'd)*

		Classifications					
		Positional			*Bag-of-words*		
Word-1	*Word-2*	Frequency	*PMI*	*PMI range*	Frequency	*PMI*	*PMI range*
red	brick	√	√	-	√	√	-
red	bus	√		-	√		-
red	houses	√		-	√		-
short	straight	√	√	-	√	√	
shot	target	√	√	-	√	√	-
sit	comfortably	√			√		
sob	throat	√		-	√		-
son	mother						
son	years						
straight	line	√			√		
stuffed	bird	√		-	√		-
sweat	face	√		-	√		
swing	arm	√			√	√	
take	houses	√					
tall	tower	√	√	√	√	√	√
tempt	person			-			-
think	idea	√		-	√		-
time	convince	√	√	-	√	√	
touch	ground	√	√	√			
ugly	face	√			√		
warm	climate	√			√		
watch	films	√					
wet	road	√		-	√	√	-
white	houses	√	√	-	√		-
white	sand	√			√		
wild	dog	√	√		√		
wind	rain	√	√	-	√	√	-
true	accurate	√	√	-	√		-
Idioms							
above	station		√	-			-
above	station			-		√	-
alive	kicking	√		-	√		-
bag	bones		√	√		√	
bat	hell	√		-	√		-
beaten	path			-			-
bird	told	√		-	√		-
bitter	end	√	√	-	√	√	√
blaze	trail	√		-	√		-

continued on next page

Table B.2: *Raw results — Cowie test data set (cont'd)*

		Classifications					
		Positional			*Bag-of-words*		
Word-1	*Word-2*	Frequency	*PMI*	*PMI range*	Frequency	*PMI*	*PMI range*
chivalry	dead	√	√	-	√	√	-
cut	dried	√		-	√		-
cut	loose	√		-	√		-
cut	pasted	√		-	√		-
darken	door	√		-	√		-
diamond	water	√		-	√		-
dog	day	√		-	√		-
dog	life	√	√	-	√	√	-
double	take	√	√	√			
double	talk	√		-	√		-
drink	fish	√		-	√		-
drive	bargain	√	√	-	√	√	√
drop	bucket	√		-	√		-
drown	sorrow	√		-	√		-
dry	dust	√		-	√	√	-
eat	dog	√	√	-			
end	road	√	√		√	√	
explore	avenue	√		-	√		-
far	cry	√	√	-	√	√	-
feather	cap	√		-	√		-
flat	board			-			
flight	fancy	√	√	-	√	√	√
flying	colours	√		-	√	√	-
fresh	daisy	√		-	√	√	-
funny	business	√			√		
gentleman	agreement	√	√	-	√	√	√
give	inch	√			√		
golden	opportunity	√	√	√	√	√	√
heart	heart	√	√	-	√		
hit	bottle	√	√	√			
jet	set	√		-	√	√	-
kill	time	√		-	√		
kiss	death	√		-	√	√	-
know	ropes	√		-	√		-
lame	duck	√		-	√		-
met	match	√					
mine	information	√	√	-	√	√	-
old	hills		√			√	
pig	ear	√		-	√		-

continued on next page

Table B.2: *Raw results — Cowie test data set (cont'd)*

		Classifications					
		Positional			*Bag-of-words*		
Word-1	*Word-2*	Frequency	*PMI*	*PMI range*	Frequency	*PMI*	*PMI range*
play	possum	√		-	√		-
pound	beat	√	√	-	√	√	-
rags	riches	√		-	√		-
rain	cats	√		-	√		-
rat	race	√		-	√		-
red	carpet	√		-	√		-
red	herring	√		-	√		-
red	tape	√	√	-	√		-
short	sweet	√	√	-			
shot	dark	√		-	√		
sit	tight	√	√	√		√	√
sob	story	√		-	√		-
son	bitch	√		-	√		-
son	gun	√		-	√		-
spit	polish	√		-	√		-
straight	answer	√	√	-	√	√	-
stuffed	shirt	√		-	√		-
sweat	blood	√		-	√	√	-
swing	cat	√	√	-	√	√	-
take	powder	√					
tall	tale	√	√		√	√	
tempt	fate	√		-	√		-
think	tank	√		-	√		-
time	tell	√			√		
touch	wood	√	√	-	√	√	
twinkling	eye	√	√	-	√	√	√
ugly	sin	√	√	-	√		-
warm	toast		√			√	
watch	clock	√					
wet	blanket	√		-	√	√	-
wheeling	dealing	√		-	√		-
whipping	boy	√		-	√		-
white	elephant	√		-	√		-
white	lie	√		-	√		-
wild	oats	√		-	√		-
wind	change	√	√	-	√	√	√
true	blue	√		-	√		

Bibliography

BNC. 2000. *Reference Guide for the British National Corpus (World Edition).* Oxford University Computing Services, http://www.natcorp.ox.ac.uk/archive/index.xml.

Thorsten Brants and Alex Franz. 2006. Web 1T 5-gram version 1. Linguistic Data Consortium. Philadelphia.

Y. Choueka, S. T. Klein, and E Neuwitz. 1983. Automatic retrieval of frequent idiomatic and collocational expressions in large corpus. *Association for Literary and Linguistic Computing Journal*, 4(1):34–38.

Kenneth Church, William Gale, Patrick Hanks, and Donald Hindle. 1991a. Using statistics in lexical analysis. In Uri Zernik, editor, *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, pages 115–164. Lawrence Erlbaum.

Kenneth Church, William Gale, Patrick Hanks, and Donald Hindle. 1991b. Using statistics in lexical analysis. In Uri Zernik, editor, *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*. Lawrence Erlbaum.

Kenneth W. Church and P. Hanks. 1989. Word association norms, mutual information and lexicography. In *Proceedings of the 27th Annual Conference of the Association for Computational Linguistics*, pages 76–83. New Brunswick, NJ.

Anthony P. Cowie, Ronald Mackin, and Isabel R. McCaig. 1983. *Oxford Dictionary of Current Idiomatic English*, volume 2. Oxford University Press.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

Afsaneh Fazly. 2007. *Automatic Acquisition of Lexical Knowledge about Multiword Predicates*. Ph.D. thesis, Department of Computer Science, University of Toronto.

Christiane Fellbaum, editor. 1998. *Wordnet: an electronic lexical database*. The MIT Press, Cambridge, Massachusetts and London, England.

John R. Firth. 1957. A synopsis of linguistic theory 1930–1955. In F. R. Palmer, editor, *Selected Papers of J. R. Firth 1952–1959*, pages 1–32. London: Longman, 1968, Phililogical Society. Oxford.

Vasileios Hatzivassiloglou and Kathleen R. McKeown. 1997. Predicting the semantic orientation of adjectives. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 174–181. New Brunswick, NJ.

John S. Justeson and Slava M. Katz. 1995. Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1(1):9–27.

Graham Katz and Eugenie Giesbrecht. 2006. Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 12–19. Sydney.

Dekang Lin. 1998a. Automatic retrieval and clustering of similar words. In *COLING-ACL '98: 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistic*, pages 768–774. Montreal, Canada.

Dekang Lin. 1998b. Dependency-based evaluation of MINIPAR. In *Workshop on the Evaluation of Parsing Systems*. Granada, Spain.

Dekang Lin. 1999. Automatic identification of non-compositional phrases. In *37th Annual Meeting of the Association for Computational Linguistics*, pages 317–324. University of Maryland, College Park, Maryland.

Christopher D. Manning and Hinrich Schütze. 2000. *Foundations of Statistical Natural Language Processing*. The MIT Press.

Kathleen R. McKeown and Dragomir R. Radev. 2000. Collocations. In Robert Dale, Hermann Moisl, and Harold Somers, editors, *A Handbook of Natural Language Processing*. Marcel Dekker.

I. Dan Melamed. 1997. Automatic discovery of non-compositional compounds in parallel data. In *2nd Conference on Empirical Methods in Natural Language Processing (EMNLP'97*, pages 97–108.

Rosamund Moon. 1998. *Fixed expressions and idioms in English: a corpus-based approach*. Oxford studies in lexicography and lexicology. Oxford,New York,Clarendon Press.

David S. Moore and George P. McCabe. 1989. *Introduction to the practice of statistics*. Freeman, New York.

Geoffrey Nunberg, Ivan A. Sag, and Thomas Wasow. 1994. Idioms. *Language*, 70(3):491–538.

Darren Pearce. 2001. Synonymy in collocation extraction. In *The NAACL '01Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations*. NAACL, Pittsburgh, PA.

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James Pennebaker. 2006. Effects of age and gender on blogging. In *Proceedings of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*, pages 199–205.

Frank Smadja. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics*, 7(4):143–177.

Dominic Widdows and Beate Dorow. 2005. Automatic extraction of idioms using graph analysis and asymmetric lexicosyntactic patterns. In *Proceedings of the ACL-SIGLEX Workshop on Deep Lexical Acquisition*, pages 48–56.