

# Synonymy in Collocation Extraction

Darren Pearce

School of Cognitive and Computing Sciences (COGS)

University of Sussex

Falmer

Brighton

BN1 9QH

UK

## Abstract

This paper describes the use of WordNet in a new technique for collocation extraction. The approach is based on restrictions on the possible substitutions for synonyms within candidate phrases. Following a general discussion of collocations and their applications, current extraction methods are briefly described. This is followed by a detailed description of the new approach and results and evaluation of experiments that utilise WordNet as a source of synonymic information.

## 1 Introduction

As an example of a collocation, Lin (1998) explains that even though *baggage* and *luggage* are synonyms, only *baggage* can be modified by *emotional*, *historical* or *psychological*.

This lack of valid substitution for a synonym is a characteristic of collocations in general (Manning and Schütze, 1999). This paper describes the formulation of a new approach to collocation extraction that is based on this observation and the results and evaluation of experiments that use it.

The definition of the exact nature of a collocation varies from one researcher to the next. It is variously defined as a habitual word combination (Lin, 1998) or a recurrent word combination (Gitsaki et al., 2000). More specifically, Smadja (1993) identifies four characteristics of collocations that have implications for machine applications, namely that collocations are arbitrary, domain independent, recurrent and cohesive lexical clusters.

## 2 Motivation

### 2.1 Foreign Language Learners

In the last few years, there has been increased interest in the English as a Foreign Language (EFL) literature in collocational information (Gitsaki et al., 2000). Howarth and Nesi (1996) also approach the use of collocations from the foreign language learner perspective. In their introduction, they cite Allerton (1984):

So often the patient language-learner is

told by the native speaker that a particular sentence is perfectly good English ... but that native speakers would never use it. How are we to explain such a state of affairs?

This is one of the questions that extraction of collocations by any reliable technique could help answer. The approach described in this paper, however, is in a position to answer this question more fully due to the extraction of ‘anti-collocations’ (see Section 5).

Gitsaki et al. (2000) give several examples of problematic word combinations in English:

<i>many thanks</i>	<i>*several thanks</i>
<i>strong coffee</i>	<i>*powerful coffee</i>
<i>tap water</i>	<i>*pipe water</i>

This last example differs from the others in that *many/several* and *strong/powerful* are confusions between synonyms (or near-synonyms). In contrast, *tap/pipe* is a confusion of a different kind since it mixes the generic and the specific. It is observations such as these that lend weight to the approach presented in this paper. The possibility of generalising from synonyms to semantic similarity is discussed as part of future research in Section 10.

### 2.2 Natural Language Processing

Collocational information is useful in a wide range of natural language processing contexts. However, it is especially relevant for some such as natural language generation, machine translation and text simplification.

**Natural Language Generation (NLG)** Generating text from a logical form requires knowledge about valid combinations of words. Manning and Schütze (1999) state that collocations are characterized by limited semantic compositionality and Howarth and Nesi (1996) claim that most sentences contain at least one collocation. The combination of these two observations carries severe implications for an NLG task working without such information:

- Word sequences may be generated that constitute a collocation accidentally. According to Manning and Schütze (1999), collocations usually have an added element of meaning. This linguistic effect may not be what was intended and could significantly change the quality of the generated text;
- Given that collocations occur so frequently, ignoring such informationally rich lexical combinations could lead to over-complicated text.

**Machine Translation (MT)** According to Gitsaki et al. (2000), collocations differ from language to language. For example, a *clear road* in English is a *free road* in Greek. Similarly, a *heavy drinker* in English is a *strong glass* in Greek.

The discrepancy that exists between mappings of collocations between different languages is problematic not only for foreign language learners but also for computational tasks such as machine translation.

**Automatic Simplification** Access to collocational information is also crucial in a text simplification setting. An example of such a task is the PSET<sup>1</sup> project (Carroll et al., 1999).

Part of the PSET implementation involves replacing difficult words with simpler alternatives. The choice of words is determined through reference to WordNet (Miller, 1990) and difficulty is measured according to the Kucera-Francis (KF) frequency obtained from the Oxford Psycholinguistic Database (Quinlan, 1992).

Simplification taking place in such a way without knowledge of collocational constraints can lead to ill-formed and awkward text. This drastically reduces the effectiveness of the simplification process.

For example, the synonyms *study* and *report* receive a KF frequency of 246 and 174 respectively. Simplifying the phrase *end of year report* to *end of year study* leads to an unnatural reading.

### 3 Existing Extraction Techniques

Although collocations usually have a strong syntactic flavour, some of the first attempts at automatic extraction from text aimed to find  $N$ -grams from 2 to 6 words in length (Choueka, 1988). However, many collocations can involve non-adjacent words. For example, *I break down doors*, *I broke down the door* and *I broke down the battered, old door* all contain the collocation  $\langle \textit{break-down}, \textit{door} \rangle$ <sup>2</sup> but with a varying number of words between the collocates.

Church and Hanks (1990) describe techniques that used mutual information to measure the strength of association between words. Potentially, this could

be used directly for collocation extraction. However, this leads to some strange ‘collocations’ such as  $\langle \textit{doctor}, \textit{hospital} \rangle$ . Just because words occur together frequently does not mean they form a collocation.

The approach taken by Smadja (1993) to overcome this problem is to *infer* syntax by measuring the spread of the distribution of counts between the two collocates. The crucial intuition here is that if there is a narrow, peaked spread then this is an indication of a syntactic relation between the two words.

Smadja (1993) also details techniques for the extraction of collocations of arbitrary length.

With the increased availability of wide-coverage grammars, current approaches favour explicit syntactic information using shallow-parsing techniques as opposed to the implicit syntax strategy of Smadja (1993). This is exemplified by Lin (1998) who bases his extraction method on dependency triples obtained from a shallow-parsed text corpus.

Although not using a parser explicitly, Justeson and Katz (1995) use patterns of parts-of-speech to extract technical terms (which are closely related to collocations). This is carried out after a thorough survey of their common syntactic variants.

## 4 A New Approach

The example of  $\langle \textit{emotional}, \textit{baggage} \rangle$  given in Section 1 consists of two words which would form a collocation according to most definitions (Choueka, 1988; Smadja, 1993; Lin, 1998). However, the observation about restrictions on possible substitution for synonyms leads to a new definition of a two-word collocation:

**Definition:** A pair of words is considered a collocation if one of the words significantly prefers a particular lexical realisation of the concept the other represents.

So, *emotional* significantly prefers *baggage* over *luggage* and similarly for *historical* and *psychological*.

This new definition contrasts to others in that there is an inherent directionality. This is henceforth referred to as *collocation preference*.

It is important to note that the definition as given above does not preclude the possibility of both words expressing a collocational preference on the other. This leads to a potential new categorisation of collocations based on the direction and number of collocational preferences. This will be developed in future research.

### 4.1 Resources Required

The majority of existing collocation extraction techniques do not take advantage of the wide range of lexical resources that are available. This contrasts

<sup>1</sup>PSET: Practical Simplification of English Text.

<sup>2</sup>Collocations will be typeset using angled brackets in this way for the rest of this paper.

with the new technique proposed here since it necessitates the availability of a map from a given word to its synonyms for each of its senses. One resource that lends itself particularly well to this task is WordNet (Miller, 1990). It is this resource that is used in the experiments.

## 5 Anti-collocations

Using sets of synonyms in collocation extraction has added benefits. With respect to a particular target word, it is possible to partition a synonym set into three disjoint subsets:

- those words which are collocations of the target word;
- those words which tend not to be used with the target word although, if used, do not lead to unnatural readings;
- those words which must *not* be used with the target word since they will lead to unnatural readings.

This last subset has been named *anti-collocations*. This type of malformed linguistic construction partly addresses the question posed in the quote by Allerton (1984) given in Section 2. As Gitsaki et al. (2000) explain, joining words that are semantically compatible does not always produce an acceptable combination. This is one reason why such information would be useful. Figure 1 gives some examples of collocations and anti-collocations.

<i>emotional</i>	{ <i>baggage</i> } { * <i>luggage</i> }	{ <i>many</i> } { * <i>several</i> }	<i>thanks</i>
<i>very</i>	{ <i>tasty</i> } { * <i>delicious</i> }	{ <i>strong</i> } { * <i>powerful</i> }	<i>coffee</i>
<i>in my</i>	{ <i>opinion</i> } { * <i>point of view</i> }	{ <i>dark</i> } { * <i>darkness</i> }	<i>room</i>

Figure 1: Examples of collocations and anti-collocations. Examples were taken from Gitsaki et al. (2000), Lin (1998) and the author’s personal interactions with foreign speakers of English.

The utility of extracting such malformed constructions is particularly useful for foreign language learners and directly relevant to natural language generation applications.

## 6 A Worked Example

To reinforce the intuitions underlying the new approach, this section extends the example of *baggage* and *luggage* by semi-automatically determining several collocations of each word.<sup>3</sup>

<sup>3</sup>The two words comprise a synset in WordNet 1.6.

1. Two million parsed sentences (50 million words) of the BNC were searched for occurrences of non-clausal modification of *baggage* and *luggage*. This data was obtained from modifier grammatical relations produced by Carroll and Briscoe’s robust statistical parsing system (Carroll et al., 1998). If the difference between their occurrence counts with respect to a particular word was *at least two* then this was deemed to be sufficient to consider them a (potential) collocation for use in subsequent stages. Counts for morphological variants such as *rack* and *racks* were manually combined.
2. For each of the bigrams extracted from the previous stage, AltaVista’s advanced search was used to obtain estimates of the occurrence counts of these phrases on the World Wide Web. Morphological variation was *not* taken into account.<sup>4</sup>
3. Details of collocations according to the Cambridge International Dictionary of English (CIDE) (Procter, 1995) were obtained. This information was used as a ‘standard’ by which to judge the application of this new technique to this particular example.

Word	BNC		AltaVista		CIDE	
<i>allowance</i>	5	0	<b>3279</b>	502	1	0
<i>area</i>	3	1	<b>1814</b>	1434	-	-
<i>car</i>	4	0	<b>3324</b>	357	1	0
<i>carts</i>	2	0	806	<b>1268</b>	-	-
<i>compartment</i>	1	3	2890	<b>5144</b>	0	1
<i>hall</i>	2	0	197	41	-	-
<i>handler</i>	5	0	<b>1448</b>	83	1	0
<i>label</i>	0	6	103	<b>333</b>	0	1
<i>Laura</i>	0	2	0	5	-	-
<i>mules</i>	2	0	30	4	-	-
<i>rack</i>	0	8	164	<b>14773</b>	0	1
<i>room</i>	3	0	927	<b>958</b>	-	-
<i>tag</i>	-	-	597	<b>3320</b>	0	1
<i>train</i>	5	0	<b>804</b>	51	-	-
<i>trolley</i>	3	0	123	<b>313</b>	-	-
<i>van</i>	0	2	190	<b>710</b>	0	1

Table 1: Counts of collocates of *baggage* and *luggage* in three sources: the BNC, AltaVista and the CIDE. Each pair of numbers consists of two co-occurrence counts of the word in the first column. The first of this pair is co-occurrence with *baggage*, the second with *luggage*. The **higher frequency** is shown in bold, the **lower frequency** in grey. Dashes represent zero co-occurrence counts with both *baggage* and *luggage*.

<sup>4</sup>This search took place on September 5th 2000.

AltaVista	CIDE	Classification
✓	✓	collocation
✓	✗	wrong
✓	A	potential
✗	✓	wrong
✗	✗	wrong
✗	A	unknown

Table 2: Classification of potential collocations. A ‘✓’ indicates agreement with the BNC frequencies, a ‘✗’ indicates disagreement and ‘A’ indicates that information was unavailable. ‘wrong’ is assigned when the CIDE contradicts the BNC, AltaVista or *both* the BNC and AltaVista.

collocation	potential	unknown
<i>baggage allowance</i>	<i>baggage area</i>	<i>carts</i>
<i>baggage car</i>	<i>baggage hall</i>	<i>room</i>
<i>luggage compartment</i>	<i>luggage Laura</i>	<i>trolley</i>
<i>baggage handler</i>	<i>baggage mules</i>	
<i>luggage label</i>	<i>luggage train</i>	
<i>luggage rack</i>		
<i>luggage van</i>		

Table 3: Categorisation of pairs derived from the BNC according to the categories in Table 2.

## 6.1 Discussion

The data obtained from this three-stage process is shown in Table 1. This frequency information was used to classify potential collocations into four categories: collocation, potential, unknown and wrong. This category was determined based on the agreement of AltaVista and the CIDE with the BNC. Table 2 shows the possible combinations of agreement and the corresponding categories.

Table 3 shows the results of this classification. The first point to note is that the table does not require a *wrong* column. This is encouraging. The data also show that except for *luggage tag*, the technique used on the BNC data obtains all the collocations listed in the CIDE and that AltaVista frequencies compare well.

Interestingly, almost as many potentially new collocations are also extracted. With the exception of *luggage Laura* (which is due to noise<sup>5</sup>) these all seem sensible. Those pairs classified as unknown also have closer AltaVista page counts in comparison to the rest of the data.

<sup>5</sup>This is actually due to AltaVista ignoring sentence boundaries when indexing. It refers to an article written by Laura Pulfer of The Cincinnati Enquirer (12/7/97) about the ‘carry-on luggage syndrome’.

## 7 Formalisation

The formalisation in this section assumes as input a sequence of (possibly ordered) pairs of words,  $p^1 \dots p^N$ . These pairs could come from a variety of sources such as parser dependencies or word bi-grams. For simplicity, it is assumed that the words in each pair are not sense-tagged or part-of-speech tagged. It is, however, straightforward to adapt the following formalisation to handle richer contexts.

The occurrence count,  $c(w_a, w_b)$ , of a particular pair of words  $\langle w_a, w_b \rangle$  is defined by:

$$c(w_a, w_b) = \sum_{i=1}^N \delta(p^i = \langle w_a, w_b \rangle)$$

where  $\delta(x)$  is 1 if  $x$  is true and 0 if  $x$  is false. WordNet is defined as a set of synsets,  $\mathcal{W}$ , where

$$\mathcal{W} = \{\mathcal{S}_1, \mathcal{S}_2, \dots\}$$

Each synset consists of a set of words which realize the same concept. It is necessary to obtain the set of synsets where, for a given word, at least two elements in the synset have non-zero co-occurrence counts.<sup>6</sup> The co-occurrence set,  $cs_w$ , of a word,  $w$ , is defined as:<sup>7</sup>

$$cs_w = \{w_v : c(w, w_v) > 0\}$$

Synsets are filtered with respect to  $w$  to obtain its *Candidate Collocation Synsets*,  $CCS_w$ , where  $CCS \subseteq \mathcal{W}$  is defined by:<sup>8</sup>

$$CCS_w = \{\mathcal{S} \in \mathcal{W} : |\mathcal{S} \cap cs_w| > 1\}$$

Thus, each candidate collocation synset consists of at least two elements whose co-occurrence count with  $w$  is non-zero.

For a synset,  $\mathcal{S} \in CCS_w$ , a word  $w'$  is selected as the most frequently co-occurring element with the word  $w$ . Its corresponding frequency is  $f'$ :

$$w' = \arg \max_{w_v \in \mathcal{S}} c(w, w_v)$$

$$f' = \max_{w_v \in \mathcal{S}} c(w, w_v)$$

The highest co-occurrence frequency,  $f''$ , of the remaining words in the synset is then calculated:

$$f'' = \max_{w_v \in \mathcal{S}'} c(w, w_v) \text{ where } \mathcal{S}' = \mathcal{S} - w'$$

It is the difference between the occurrence counts of these two top-ranked elements,  $f' - f''$ , that can be

<sup>6</sup>This varies slightly from the treatment of the *baggage* and *luggage* data given in Section 6. This is because zero co-occurrence counts would occur very frequently within any given synset.

<sup>7</sup>A similar definition exists for collocation preference in the opposite direction.

<sup>8</sup>In reality, a more efficient approach is taken since it is not necessary to process every synset in WordNet.

used to rate ‘collocation strength’,  $s$ , in the following way:

$$s = \frac{f' - f''}{f'}$$

Division by  $f'$  ensures that  $0 \leq s < 1$ . A value of  $s \approx 1$  indicates high collocation strength and  $s \approx 0$  indicates low.

## 8 Experiments

Approximately 50 million words of parsed BNC data were used with the approach described above. This was the same data as in Section 6 (Carroll et al., 1998). Accompanying each of the parsed sentences was a list of corresponding non-clausal modifications. These were then used as the input pairs to the formalisation detailed in Section 7.

Each word pair was tagged with its corresponding part-of-speech. However, these were not used in the experiments although future work will utilise this information. No morphological processing was applied to the input so *limit* and *limits* were treated as different words.

The data resulting from the experiment includes phrases such as:

<i>human beings</i>	<i>human rights</i>
<i>living room</i>	<i>ground floor</i>
<i>human nature</i>	<i>welfare state</i>
<i>education system</i>	<i>bedroom door</i>
<i>managing director</i>	<i>labour party</i>

However, several high-scoring phrases are repeatedly extracted based on counts in the wrong synsets. These phrases look like collocations but are in fact extracted by coincidence.

### 8.1 Coincidental ‘Collocations’

Coincidental ‘collocations’ occur in this approach due to lack of semantic information. Figure 2 shows some examples of these. Each braced group represents the elements of a synset whose co-occurrence counts with the target word is non-zero. The element in bold indicates the word with the highest co-occurrence frequency.

All of the phrases in Figure 2 are collocations. However, the reasons for their extraction is coincidence. The word *post* in *post office* gets high counts for three synsets none of which represents the postal service (the correct sense).

In the cases of *upper limit* and *speed limits*, the justifying synset corresponds to a ‘central nervous system stimulant’. These coincidences also mean that *upper limits* and *speed limit* are *not* extracted.<sup>9</sup>

<sup>9</sup>Combining counts for morphological variants would go some way to improving this situation but would not eliminate the problem.

Collocation	Justifying Synsets
<i>last year</i>	$\left\{ \begin{array}{l} \textit{finish} \\ \textbf{last} \\ \textit{close} \end{array} \right\}$ $\left\{ \begin{array}{l} \textit{death} \\ \textbf{last} \end{array} \right\}$ $\left\{ \begin{array}{l} \textit{end} \\ \textbf{last} \end{array} \right\}$
<i>post office</i>	$\left\{ \begin{array}{l} \textbf{post} \\ \textit{station} \end{array} \right\}$ $\left\{ \begin{array}{l} \textbf{post} \\ \textit{office} \end{array} \right\}$
<i>upper limit</i>	$\left\{ \begin{array}{l} \textbf{upper} \\ \textit{speed} \end{array} \right\}$
<i>speed limits</i>	$\left\{ \begin{array}{l} \textit{upper} \\ \textbf{speed} \end{array} \right\}$

Figure 2: Some examples of repeatedly extracted coincidental ‘collocations’.

## 9 Evaluating Coincidence Rate

It is useful to determine what proportion of erroneous collocations are accounted for by coincidence. This section describes an experiment to evaluate this quantity by using the semantic concordance (semcor) that accompanies the WordNet 1.6 distribution.

Collocations were extracted from a parsed version of semcor<sup>10</sup> in two ways: using the sense information and ignoring it. These sets of data were then compared.

Figure 3 shows the number of collocations extracted across varying strengths for both runs of the algorithm. In general, coincidence accounts for about 80% of the collocations extracted without using sense information.

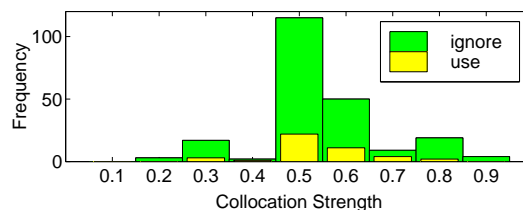


Figure 3: Evaluating coincidence rate using semcor. Collocation strengths were rounded down to the nearest 0.1.

## 10 Conclusions and Future Work

This paper has presented a new technique for collocation extraction which relies crucially on a mapping from a word to its synonyms for each of its senses. WordNet is used to fulfil this function.

<sup>10</sup>Obtained using the work of Carroll et al. (1998).

The technique proposed looks promising. There are also many ways in which this technique is currently being developed.

One of the most important of these is the utilisation of the rich hierarchical structure of WordNet. This includes lexical relations such as hypernymy and meronymy. This information will be integrated into the approach by generalising from the idea of synonymy to conceptual similarity. This will require a sound probabilistic backbone.

Implicit in the formulation given in this paper is the assumption that any given synset (or conceptually related group of words) has one and only one element that forms a collocation with a particular target word. This may not be the case as is suggested by the discussion of anti-collocations in Section 5. Such situations could also be accounted for by a probabilistic approach.

Evaluation of collocation extraction methods is difficult and one that Smadja (1993) feels is best performed by a lexicographer. It is expected that the BBI Combinatory Dictionary of English (Benson et al., 1986) will be used (among other Machine Readable Dictionaries) as a standard by which to judge this new technique and the variants that will be developed.

## Acknowledgements

This work is being carried out under a studentship attached to the project 'PSET: Practical Simplification of English Text' funded by the UK EPSRC (ref GR/L53175). Further information about PSET is available at <http://osiris.sunderland.ac.uk/~pset/welcome.html>.

I would also like to thank my supervisor, John Carroll, for his continued support and advice.

## References

- David J. Allerton. 1984. Three (or four) levels of word co-occurrence restriction. *Lingua*, 63:17–40.
- Morton Benson, Evelyn Benson, and Robert Ilson. 1986. *The BBI Combinatory Dictionary of English*. John Benjamins Publishing.
- John Carroll, Guido Minnen, and Ted Briscoe. 1998. Can subcategorisation probabilities help a statistical parser. In *6th ACL/SIGDAT Workshop on Very Large Corpora*, pages 118–126, Montreal, Canada.
- John Carroll, Guido Minnen, Darren Pearce, Yvonne Canning, Siobhan Devlin, and John Tait. 1999. Simplifying text for language-impaired readers. In *Proceedings of the 9th Conference of the European Chapter of the ACL (EACL '99)*, Bergen, Norway, June.
- Yaacov Choueka. 1988. Looking for needles in a haystack or locating interesting collocational

expressions in large textual databases. In *Proceedings of the RIAO '88 Conference on User-Oriented Content-Based Text and Image Handling*, pages 1–15, Cambridge, MA, March.

- Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29, March.
- Christina Gitsaki, Nagoya Shoka Daigaku, and Richard P. Taylor. 2000. English collocations and their place in the EFL classroom. Available at: <http://www.hum.nagoya-cu.ac.jp/~taylor/publications/collocations.html>.
- Peter Howarth and Hilary Nesi. 1996. The teaching of collocations in EAP. Technical report, University of Leeds, June. Available at <http://gillett.connect-2.co.uk/baleap/reports/gl/leeds/contents.htm>.
- 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.
- Dekang Lin. 1998. Extracting collocations from text corpora. In *First Workshop on Computational Terminology*, Montreal, Canada, August.
- Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts.
- George A. Miller. 1990. WordNet: An on-line lexical database. *International journal of lexicography*.
- Paul Procter, editor. 1995. *Cambridge International Dictionary of English (CIDE)*. Cambridge University Press.
- Philip Quinlan. 1992. *The Oxford Psycholinguistic Database*. Oxford University Press.
- Frank Smadja. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1):143–177, March.