Using the Distributional Hypothesis to Derive Cooccurrence Scores from the British National Corpus

David Hardcastle
School of Computer Science & Information Systems
Birkbeck, University of London.
ahard04@dcs.bbk.ac.uk

Abstract

In this paper I examine a number of cooccurrence-based scoring systems using the British National corpus to measure word association over wide contexts. The principal aim of this paper is to address the question of how to evaluate a given scoring system, or how to compare two scoring systems, without relying on a small list of example pairs and a 'feel' for the results. I evaluate these systems using i) a list of noun-noun pairs and ii) a simple test on aligned and misaligned sets of nouns. I also consider why noun-noun pairs are deemed appropriate for such mechanisms and explore the prospects for determining which words or lemmas will be appropriate for a distributional scoring approach. For my specific application an algorithm similar to MI-score operating across multiple windows offers the best results.

1 Introduction

1.1 Application Requirements

The application for which I require a word association scoring algorithm is a cryptic crossword clue compiler. The first stage in the design unpacks rubrics for the word for which the clue is being constructed into candidate sets of key words that could be combined to produce clues. These are then sorted on a variety of criteria, the most important being word association. My assumption is that the promotion of groups of such key words that have shared word associations will lend an idiomatic feel to the clues. Since there will be a very large number of candidate sets, the scoring algorithm must evaluate rapidly and requires high precision but not necessarily high recall. Some of the words may have low frequencies in the corpus, others may be very common. The system should favour significant n-grams, but it should not just promote pairs which happen to co-occur at word boundary level and ignore pairs with a looser lexical connection, otherwise the system will only favour bigrams.

As an example, consider this clue for DIURNAL from the Independent:[1] 'everyday source of coffee in mug'. Once a rubric has been identified (URN in DIAL = DIURNAL in this case) there are a considerable number of possible expansions depending on what synonym is given for URN, DIAL and DIURNAL. In the clue URN is defined as 'source of coffee' and DIAL as 'mug'. Other synonyms for DIAL include 'face', 'clock', 'knob', 'zodiac'. The word association scoring phase of the system should identify the association between 'coffee' and 'mug' and the lack of association between 'coffee' and 'face', 'clock', 'knob' or 'zodiac'[2].

The system requirement is therefore a measure of word association with a clear decision threshold below which candidate pairs will be rejected. In this paper I consider how to evaluate competing word association measures, and how to tune the threshold of a given measure for a particular application.

[1] The Independent, 05/08/04

[2] Using my Russian Doll algorithm (decision threshold 0.40) the scores are as follows: coffee-mug 1.00, coffee-face 0.22, coffee-clock 0.25, coffee-knob 0.20, coffee-zodiac 0.14.

1.2 Word Association and Semantic Distance

Budanitsky and Hirst (2001) examine a number of semantic distance measures based on WordNet connections using a ranked list of noun-noun pairs[3] and a dummy application. Jarmasz and Spackowicz (2003) use the same datasets to examine the use of Roget's 1987 Thesaurus to determine semantic distance.

In both of these papers there are two tests for each system: first a rank correlation against some expert-ranked data that is taken to be normative, and second use within a particular application in which success or failure can be objectively measured (to some degree)[4].

For my application semantic distance is not the right measure. I want the pairs chosen by the system to encourage the reader of the clue to build some narrative behind it based on some shared semantic context or word association, but this does not equate to proximity of definition. For example, the pair coast-forest has a low score in the normative list of pairings, but is a perfectly acceptable pair for word association. Conversely the pair crane-implement is close in semantic distance but diverse in terms of word association.

Reading the previous paragraph, you may dispute the notion that coast-forest shares a greater degree of word association than crane-implement, and I would have no normative list or definition to cite in my defence, only my own opinion. The challenge in this paper, then, is a means of comparing and using word association algorithms with neither a set of right and wrong answers, nor a succeed or fail application against which to measure them.

2 The Evaluation Method

Since there is no objective background against which to evaluate the measures I do not claim that the method which I set out in this paper determines the best word association measure, nor that it determines the correct decision threshold for each one. Rather, it assists me (the system developer) in choosing an appropriate measure, and in setting the decision threshold at a level appropriate to my application requirements. For a different application different source data may be used, and different requirements may drive the choice of decision threshold.

I apply two tests to the word association measures, the first identifies the decision threshold to use, the second checks that scores for a set of hand-picked positive and negative pairs fall into two discrete groups. It is obviously important that the data used in both tests is finalised before the evaluation is carried out.

To set the decision threshold I use aligned and misaligned cross-products of related sets of words to generate a large number of pairings that should pass or fail a word association score test. I then measure approximate precision and recall and tune the threshold to a precision and recall that is appropriate to the system. I use four pairs of sets, each pair having a set of related words and then a co-domain set, as follows:

A [types of clothes]	B [things relating to clothes]
C [types of weather]	D [things relating to weather]
E ['maths' 'mathematics' 'mathematician']	F [things relating to maths]
G ['consultant' 'consultancy']	H [things relating to consultancy]

I generate a match list from the union of the cross-products AXB, CXD, EXF and GXH, and a mismatch list from the union of the cross-products AXH, GXB, CXF and EXD. Using cross-products in this way provides me with fairly substantial match and mismatch lists without the need to consider each entry individually.

[3] The list of pairs used is a subset of Rubenstein and Goodenough's set of 65 expert-chosen pairs (Rubenstein and Goodenough 1965) selected by Miller and Charles (1991).

[4] Budanitsky and Hirst use an application that detects real-word spelling errors running against a sample for which the correct spellings are known to provide precision and recall data for the different semantic distance measures.

For the second test I constructed a list of hand-picked noun-noun pairs with the help of my partner by brainstorming in a local café[5]. There are four categories for the list, based on proximity of association with twenty pairs in each: word boundary pairs (e.g. wine-bottle), pairs from short phrases (e.g. cat-dog), associated pairs (e.g. pizza-salami) and nonsense pairs (e.g. apron-lion). In the second test I rate the systems on their ability to separate the two lists cleanly. Given the demands of my particular application, I would prefer to find a measure that did not grade the four categories in a sliding scale since pairs with loose associations are as relevant to the application as word boundary pairs.

The full list of noun-noun pairs along with their scores, and the list of words in each of the above sets are given in the Appendices. I and my partner chose these lists somewhat arbitrarily, so they do not represent a prescriptive list, however they form a workable starting point from which to make some meaningful statements about the scoring systems.

3 Word Association Scoring Methods

I evaluate a variety of word association scoring methods in this paper, as follows:

- Average association ratio
- t-score (over wide windows)
- Inverse Distance Measure
- Dice Coefficient (by paragraph and sentence)
- Russian Doll Algorithm (weighted MI-score over multiple windows)

The scoring methods which I use in this paper rely on the Distributional Hypothesis and were run against the World Edition of the BNC. By Distributional Hypothesis I mean the Firthian view that words which have similar meaning will share similar distributions: “the meaning of entities […] is related to the restriction on combinations of these entities” (Harris 1968: p12); see (Dagan n.d.). I have restricted the scope of this paper to noun-noun pairs for simplicity, but I consider ways to include other classes of words below.

Most measures based on the Distributional Hypothesis find n-grams within given (usually small) window sizes. However, Church and Hanks (1990: p4) suggest that “it might be interesting to consider alternatives … that would weight words less and less as they are separated by more and more words” although they choose a window of ±5 words for the basis of their paper. Gale, Church and Yarowsky (1993), use window sizes of ±50 words to extract contextual clues from corpora to assist in the disambiguation of polysemous nouns while Budantisky and Hirst (2001) find that a scope (window size) of between ±1 and ±2 paragraphs gives optimal performance in an application that detects malapropisms using various network-based measures of semantic distance. These window sizes would equate to approximately ±65 or ±130 words in the BNC.

This shows that different types of relationship between words operate at different window sizes, so a system that will uncover these different types of relationship must operate within multiple window sizes. For this reason all of the measures that I use in this paper operate over multiple and/or wide windows.

3.1 Average Association Ratio

I refer here to the “association ratio” (MI-score) proposed by Church and Hanks (1990: p2) to examine word pairings over a window size of ±5 words and defined as:

$I_{x,y} = \log_2 (P_{x,y} / P_x P_y)$

where $P_{x,y}$ is given by the observed matches within the window size normalized by the corpus size, and P_x and P_y are given by the frequencies of x and y normalized by the corpus size. For Church and Hanks the ordering of the match is important, so $I_{x,y}$ is not equal to $I_{y,x}$. Over larger distances word order should not be a factor and I am calculating the MI-score based on all matches within the

[5] I would have much preferred to have a list with a more objective or expert source, perhaps survey results from the Generation Game or a list compiled by a panel of linguists but I could not locate such a source.

window[6]. I calculate average association ratio for x and y occurring in multiple windows of increasing size using the following definition[7]:

$I_{x,y} = \log_2 (f_{x,y}N/ f_x f_y)$

Since the window size itself is not part of the equation this means that the score is excessive for large windows, which has a detrimental impact on precision.

I have deliberately not referred to this measure as MI-score, as the process of averaging out over large window sizes undermines the measure as proposed by Church and Hanks. It should be noted that MI-score as described in their paper performs considerably better than this experiment with it, although being restricted to narrow windows it does not pick up the looser associations that I am interested in so I have not included it in this paper.

3.2 t-score (over wide windows)

In using t-score, I again make some changes to the implementation so that I can measure data gathered from wide window sizes.

A thorough description of the t-score is given in Manning (1999: 163ff) where it is defined as:

$$t = \frac{(x - \mu)}{\sqrt{(\sigma^2/N)}}$$

The example given (by Manning) is for a bigram, so $P_x=f_x/N$ and under H_0 where the occurrences of x and y are independent $P_{x,y}=P_xP_y=f_xf_y/N^2$. The N bigrams in the corpus are considered as a Bernouilli trial in which we assign 1 to the bigram x,y and 0 to all others, with $p= P_{x,y}$ as above. The mean μ is p and σ^2 is $p(1-p)$ which is taken to equate to p for most bigrams since p is so small.

Since I am interested in multiple window sizes I calculate the probability of finding one or more matches for x and y in a window of size w. Since w is so much smaller than N I then assumed that once again we have a Bernouilli trial as above across N windows (although in practice there are only $N-w$ such windows).

3.3 Inverse Distances Measure

This measure[8] is a function of the inverse distance separating all cooccurrences within any document, summing this for the pair and then using the frequency to normalise the result. The attraction of this method is that it is not restricted to any particular window size or combination of sizes. However to determine the function to use one has to decide how the distances should be ranked and weighted. This would require answers to questions such as “should two cooccurrences 10 words apart rank higher than one of 9 and one of 12?” or “how many cooccurrences 100 words apart are equivalent to a single cooccurrence with a distance of 10 words?”.

I do not believe that cooccurrence measures (even over very large data sets) are sufficiently fine-grained to answer such questions adequately. I therefore used trial and error and settled on a particular implementation that gave some useful results. I chose the sum of the inverse squares of all of the distances normalised by the sum of the frequencies:

$$D_{x,y} = \log_{10} \frac{\Sigma(1/d_{x,y})^2 * 1000}{f_x + f_y}$$

This formula lends weight to the number of cooccurrences and more weight to the closer cooccurrences. Since the x-y matches number the same as the y-x matches I normalised by the sum of the frequencies, although this penalises pairs with very imbalanced frequencies. Lastly I chose to multiply the result by 1,000 and take the log to convert the measure to a readable scale.

Although I could not answer the questions posed above, and chose the formula through a somewhat arbitrary process, the scores assigned to a list of hand-picked pairs appear interesting, and there are

[6] Word order is of course important if collocations are found, and the system notes the ordering of these for all the scoring systems which detect word boundary collocations.

[7] This equation matches that given by Matsumoto (2003).

[8] Suggested by Trevor Fenner (personal communication)

plenty of good examples that I could provide as evidence of the success of this formula. This underlines the need for a process to evaluate and compare scoring systems, as some nice-looking hand-picked results will not guarantee that this measure will be useful within my application.

3.4 Dice Coefficient

An alternative approach to counting matches in windows is to compare the distributions of x and y using some measure of properties. I consider the list of all paragraphs in which the word x occurs in the BNC as a feature vector describing the distribution of x across the BNC. The distribution of x and y over the paragraphs of the BNC can then be determined using some statistical measure.

The Dice coefficient returns the similarity of two distributions. For the feature vectors X and Y the Dice coefficient is twice the ratio of the intersection over the union:

$$\mathrm{Dice}_{x,y} = \frac{2\ |X \cap Y|}{|X| + |Y|}$$

The features in the vectors are all of the paragraphs in the BNC, with the feature vector X recording a 1 for all of these with one or more occurrences of x and a zero for the remainder. See also Dagan, Lee and Pereira (1999) for a discussion of similarity measures and Smadja, McKeown and Hatzivassiloglou (1996) on the use of the Dice coefficient to measure cooccurrences in parallel corpora.

3.5 Russian Doll Algorithm

This system scores cooccurrences over multiple windows using weighting to take account of window size. In a previous paper (Hardcastle 2001) I took a similar approach but used the orthographic structure of the BNC to delimit the windows. In this version I use multiple word-distances within a document. This approach requires less data to be held, while using more window sizes. In an attempt to take some of the benefits of both systems I chose (where possible) window sizes that approximated to structural divisions within the BNC: ±1 (word boundary), ±5 (phrase), ±10 (sentence), ±20 (paragraph), ±50 (multi-paragraph), 100, 250, 500, 1000[9].

To avoid double-counting matches the windows are not considered to overlap, so a cooccurrence at a distance of 14 words will count towards the ±20 category and no other. I decided to score each window by taking some function of the observed matches over the expected yield, given by:

$$Ex,y = fxfyw\ /\ N$$

where w is the window size and N the number of words in the corpus, in other words the number of matches that would be expected by chance within the window assuming that occurrences of the two words (x and y) are distributed randomly across the corpus.

The assumption of independent, random distributions for x and y is controversial (see above), in my 2001 paper I proposed using a baseline set of pairings for the expected yield. This addresses the assumption of an independent, random distribution, but raised new questions, in particular how to manage pairs with imbalanced frequencies or disributions.

The headline score for the pair is the sum of the weighted scores across the windows, this gives scores in a range from 0 to many thousands. It appears from inspection that scores of 65 and over are reasonable and that scores of 500 and over are reliably positive, so I converted the scores to a 0.0-1.0 range using $\log_2$ with a ceiling of 512 (this brings the scores into a range 0-9), I then squared this and divided by 81 so that the difference between acceptability and excellence is not compressed into a small segment of the scale. Note that this transformation makes the scores more readable, but does not change the rankings of any pairs under the system. The result is as follows:

$$RDx,y = \frac{(\log_2 \max((\ \Sigma\ f_{x,y}N\ /\ f_x f_y w\ ),\ 512))^2}{81}$$

[9] Since devising this algorithm I have read Church and Gale's 1990 paper (Church and Hanks 1990) and discovered that this algorithm is pretty much calculating the MI-score for multiple concurrent windows weighted for window size.

4 Evaluating the Scoring Measures

I performed two tests on the scoring systems, as set out below.

4.1 Identifying the decision threshold

Taking the cross-product of all of the associated pairs of sets described above generates a match list of noun-noun pairs that have something in common. Taking the cross-product of other pairs of these sets generates a mismatch list of noun-noun pairs that generally have nothing in common[10]. The precision for the system is then the percentage of pairs from the mismatch list that are rejected, and the recall is the percentage of pairs from the match list accepted. I approximated thresholds for a range of precisions and compared the recalls attainable at different levels of precision.

The table below illustrates how some measures are more sustainable at high thresholds (with high precision) while for others the recall crashes when the precision is raised. For my particular application I require high precision rather than high recall.

Precision	Average Association Ratio	t-score	Inverse Distance	Dice Paragraph	Russian Doll
50.0%	92%	100%	100%	85%	92%
75.0%	86%	100%	65%	70%	85%
90.0%	70%	100%	43%	52%	69%
95.0%	55%	100%	32%	46%	58%
97.5%	38%	18%	21%	42%	43%
99.0%	21%	13%	12%	30%	26%
99.5%	16%	9%	10%	23%	22%

Table 1: Precision vs Recall for the five measures using the match and mismatch sets.

Since recall of around 10% is sufficient for the requirements of my application I chose a decision threshold for each measure that approximated a precision of 99.5%[11], giving decision thresholds for each measure as follows:

Average Association Ratio	t-score	Inverse Distance	Dice Paragraph	Dice Sentence	Russian Doll
9.00[12]	2.58	2.45	0.013	0.003	0.40

Table 2: Thresholds identified by match and mismatch sets.

4.2 Measuring Consistency and Decisiveness

Although a low recall is acceptable for my application, I don't want to miss out on loose associations and only retain pairings where there is a possible n-gram or some other sentence level association (this

[10] Although, of course, a handful of the 'mismatches' such as sky-infinity (weather x mathematics) are actually (in my view) plausible pairings.

[11] With the exception of t-score for which I used a threshold that corresponds to a confidence level of 0.005 (2.576) rather than my approximated threshold of 2.5 based on 99.5% precision.

[12] Note that this differs from Church and Hanks threshold of 3.0 for the MI-score association ratio, this is because I have averaged the ratio over multiple windows changing the measure. Interestingly, using a single window size of +-5 (as in Church and Hanks original paper) against my match/mismatch test data with a threshold of 3.0 corresponds to an approximate precision of 98%.

would make the associations in the clues too black and white). I also require the measure to be decisive; it would ideally separate matches and mismatches into distinct sets with little overlap.

To test these features of each measure, I used four groups of twenty noun-noun pairs the first three of which represent positive pairs which are hand-picked for associativity, and the fourth representing nonsense pairs. Given that the pairs have been selected by hand precision should be very high, and there should also be a minimal overlap between the positive pairs and the negative pairs, that is there should ideally be no negative pairs scoring more than the minimum score for the positives, and no positives scoring less than the maximum score for the negatives. Also, while it would make sense that the set chosen to associate at the tightest distance would score most, the recall should not drop too sharply between the first and the third set.

	Average Association Ratio	**t-score**	**Inverse Distance**	**Dice Paragraph**	**Russian Doll**
Precision	100%	100%	100%	100%	100%
Recall 1 (word boundary)	35%	100%	100%	50%	95%
Recall 2 (phrase level)	25%	100%	95%	80%	100%
Recall 3 (loose association)	5%	60%	40%	40%	100%

Table 3: Precision and Recall: over 4 hand-picked sets of 20 pairings using the threshold identified in test 1. All measures have a precision of 100% in this artificial environment. Note though that different measures are more or less resistant to wider scope associations between the pairs.

	Average Association Ratio	**t-score**	**Inverse Distance**	**Dice Paragraph**	**Russian Doll**
Overlap	14%	29%	29%	4%	0%

Table 4: Overlap: the number of negatives scoring more than the minimum positive pairing plus the number of positives scoring less than the maximum negative as a percentage of the total pairs.

All of the measures performed as expected with regard to precision, only one of the 20 hand-picked nonsense pairs scored over the threshold for any measure. There was a significant difference between the measures dependency on tight word context, even though all were designed to incorporate information from wide windows, and (accordingly) there was a wide variation in overlap (although some of this can be attributed to positive pairs scoring zero).

5 Analysis

In the first test I plotted approximated recall for increasing precision using lists of match and mismatch pairs generated from sets. I built the sets myself, but it would be possible to identify sets of related words and their co-domains using texts with tight semantic context, such as a newspaper match report, or a property feature in a magazine, for example. Generating the lists from the sets gave me a reasonable amount of data (over 700 pairs in each list) without having to think them all up.

All of the measures experience a fall in recall as precision increases (as one would expect) although the gradients are steeper. Most of the measures follow a similar profile, with the exception of the t-score where there are no meaningful results for precision of less than 95%, however some of the measures are more tolerant of a rise in precision, possible implying that they are more robust. Since precision is my principal concern I chose maximum precision[13] for all of the measures and set the thresholds accordingly.

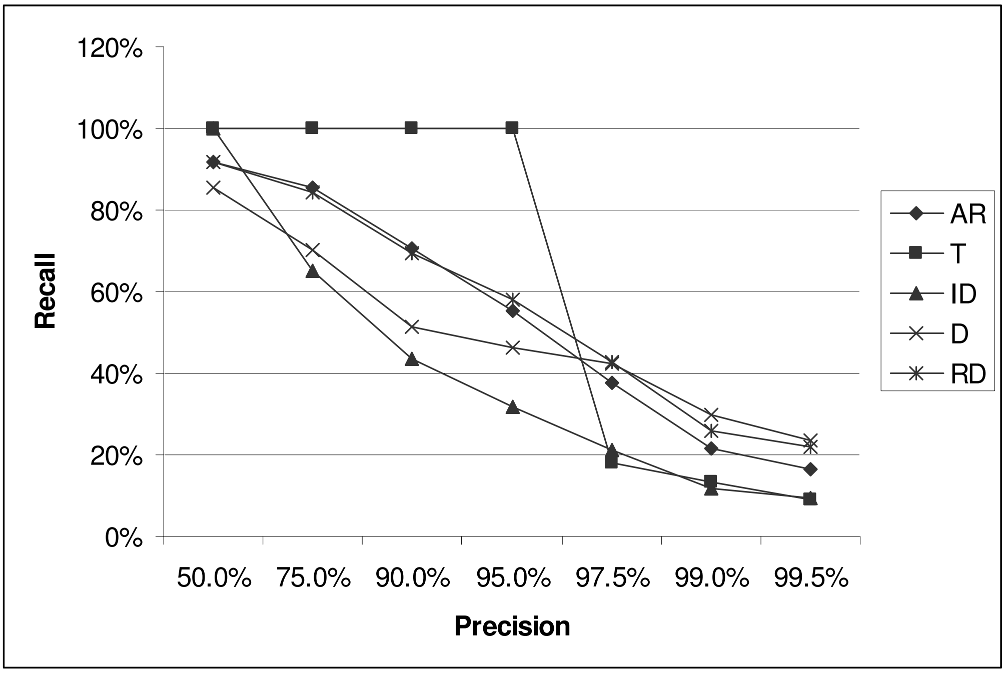

Fig 1: Recall vs Precision for the various measures.

In the second test I used the threshold to confirm a very high precision against some hand-picked nonsense pairs (all of the measures had a precision of 100%), and then to examine the recall against three lists of hand-picked pairs with looser and looser associations. Most of the measures experienced a loss of recall for the pairs with looser associations, showing that they had retained their bias toward n-grams despite my attempts to introduce information from further distances and wider windows.

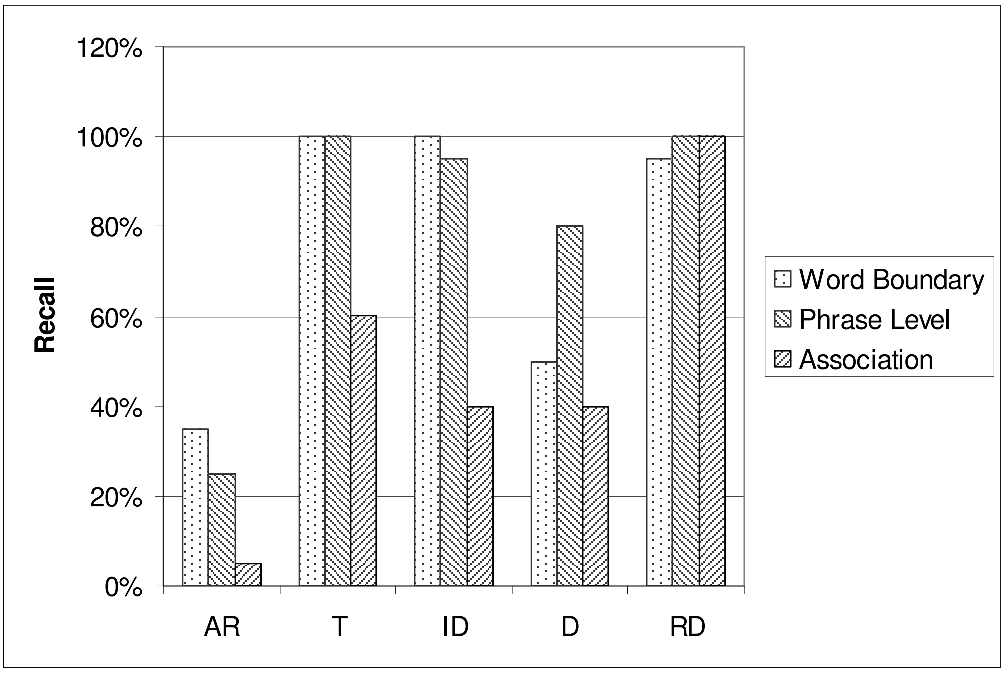

Fig 2: For most measures, recall falls as the scope of association increases.

[13] Since I only 759 mismatch pairs I cannot reasonably determine a threshold for precision of over 99.5%, although I can approximate the threshold for 99.5% from the results.

I then measured the overlap between scores for the positive pairs and for the nonsense pairs – ideally the system should be decisive and there should be no overlap between the scores for hand-picked pairs that associate and scores for hand-picked nonsense pairs. Only the Russian Doll measure had no overlap at all, this means that the scores could be divided into two discrete sets, as shown in the following graphs.

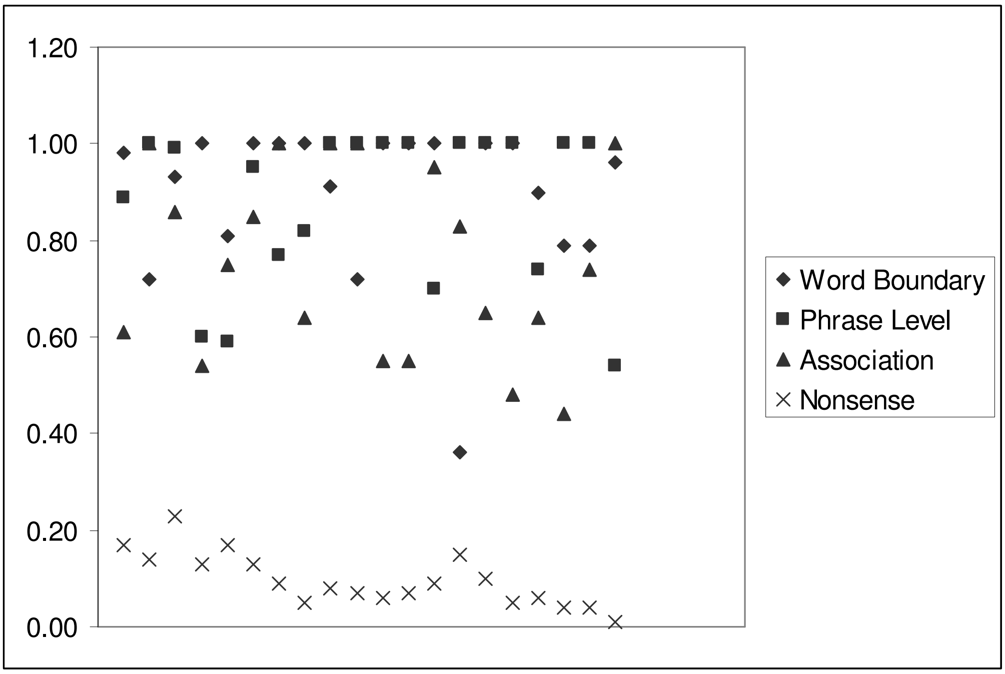

Fig 3: Scores for positive pairings and nonsense pairings using the Russian Doll algorithm fall into two discrete sets (overlap is 0%).

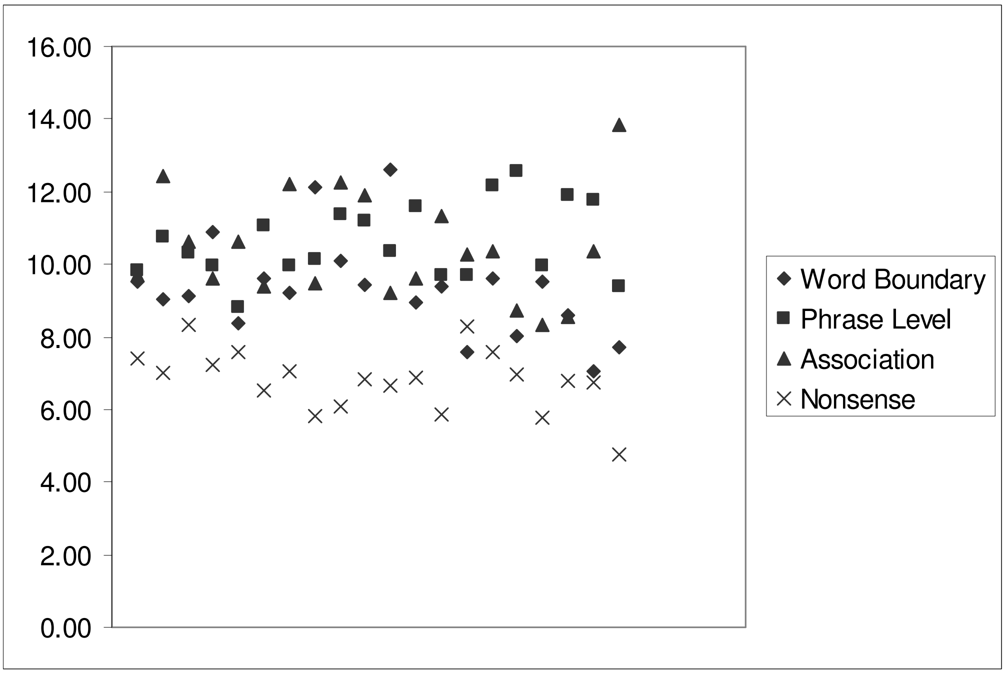

Fig 4: Scores for the positive and nonsense pairings using the average association ratio overlap leaving no clear dividing line (overlap is 14%).

5.1 Average association ratio

Initially, I chose a threshold of 3.0 for the average association score, since that is the cut-off reported by Church and Hanks for MI-score[14]. However in their paper the window size was always ±5, whereas I am using windows up to ±1000. Tuning the threshold using the second test I found that a threshold of 9.0 raised the precision to over 99% and loweeds the recall to 16%, a much more suitable combination for my application, although with the threshold at 9.0 fifteen of the hand-picked pairs would be rejected which I think is too many. Simply averaging the MI-score across the different window sizes appears to produce coarse results which cannot be tuned for the application.

5.2 T-score

Hunston differentiates between the MI-score and the t-score as follows: "MI-score is a measure of strength of collocation; t-score is a measure of certainty of collocation" (2002: p73). The t-score has a very high precision, and returns a score of zero for all of the negative pairs, which is impressive. However the recall is very low[15], and when I looked through the detail at window level I found that the t-score was negative for every single pair for all window sizes over ±10, this means that the pairs pizza-olive, dog-rabbit and cook-restaurant all have a t-score of 0.00, despite being hand-picked as associative matches. T-score is a decisive measure of n-grams but my attempt to apply it to wide windows is not useful for my application.

5.3 Inverse Distance

The inverse distance method appears to work very well when perusing the hand-picked pairs by eye. However the overlap is very high, and this is reflected in a low precision. To raise the precision to over 99% the threshold would need to be 2.4, this would mean rejecting 18 of the hand-picked pairs and would lower the recall to 10%. I made a lot of arbitrary decisions in formulating this measure, as a result it is not robust and cannot be tuned for my application.

5.4 Dice Coefficient

This measure has a low overlap, making it a decisive system, and continues to function with the threshold set for high precision. It appears to be a useful measure of word association, although it cannot identify the context of association (for example that hub-cap is actually an acceptable collocation), and does not perform so well at scoring loose associations (e.g. dog-hamster, pizza-salad).

5.5 Russian Doll Algorithm

This algorithm has a low overlap, and functions well when tuned for a precision of 99.5%. An additional attraction is that there is considerable agreement between the word boundary, phrase and broad context positive lists, that is to say that words which are found to cooccur at word boundary level do not automatically out-score words which associate within wider windows. Take for example the pairings in the following table, pairs with loose associations such as pizza-olive or cage-monkey can still score as strongly as n-gram pairs such as bus-lane or mouse-mat, whereas in other systems information from wide contexts does not have the strength to compete with local coocurrences.

	Average Association Ratio	t-score	Inverse Distance	Dice Paragraph	Russian Doll
[Threshold]	**9.00**	**2.58**	**2.45**	**0.013**	**0.40**
pizza-olive	10.36	0.00	0.00	0.002	0.65
cage-monkey	10.36	1.54	1.65	0.005	0.74
bus-lane	9.43	36.02	4.35	0.012	0.72
mouse-mat	7.06	23.36	2.71	0.003	0.79

[14] "As a very rough rule of thumb we have observed that pairs with $I(x,y)>3$ tend to be interesting, and pairs with smaller $I(x,y)$ are generally not." (Church and Hanks 1990: p5)

[15] I chose a threshold confidence level of $\alpha=0.05$ and a value for t of 2.576 as described in Manning (1999: p164)

Table 5: Pairs with loose associations have similar scores to n-gram pairs when using the Russian Doll algorithm or the average association ratio.

Additionally this algorithm could attempt to characterise the nature of the association between a pair. The following table shows how the proportion of the Russian Doll score is distributed across the different window sizes for these different categories:

	±1	**±5**	**±10, ±20**	**±50 … ±1000**
Word Boundary[16]	**71%**	6%	10%	14%
Clause	2%	**54%**	24%	21%
Context	5%	**27%**	**33%**	**35%**
Nonsense	0%	10%	14%	**76%**

Table 6: An analysis of the proportion of the score derived from each window size for each category suggests that the algorithm could attempt to characterise the nature of an association.

5.6 Conclusion

In this paper I explore a number of word association scoring systems using wide window sizes to attempt to find a measure that will recognise loose associations as well as n-grams and common phrases.

It is difficult to evaluate and compare such scoring systems, and also to determine the threshold for accepting or rejecting a pair without resorting to an arbitrary ‘feel’ for the results. I use cross products of sets to derive match and mismatch lists against which I can tune the thresholds to the precision and recall required by my application. I then test these thresholds against some hand-picked pairs (ideally these would be some normative or prescriptive list) to decide which measure is most appropriate for my application.

This approach is set out as a starting point for evaluating, comparing and tuning scoring systems for an application that identifies noun-noun pairs that are associated with one another.

The list of noun-noun pairs is perhaps rather short, and is certainly not a representative sample of British English but rather an arbitrary list that to some extent reflects the location in which it was created (an Italian restaurant). The same criticism applies to the sets which I used to build the match and mismatch list. On the other hand the approach forces all of the systems to be evaluated under the same terms and over the same data, and provides some justification for determining the threshold and for deciding which measure to use for a given task.

[16] Figures for word boundary do not include wine-bottle, coffee-cup and tea-cup all of which have a strong bias towards clause level, I think that this is due to the phrases “bottle of wine”, “cup of coffee” and “cup of tea” and shows that while I thought of these pairs as word boundary pairs they are actually more likely to be found in phrases.

6 Criticisms

In this section I raise criticisms of the scoring systems set out in this paper and consider how they might be addressed.

6.1 Dependency/Probability

Most of the approaches which I explore in this paper measure the difference between the measured distribution of x and y in the corpus with some assumed baseline behaviour for a pair of independent, randomly distributed variables. Dunning (1993) takes issue with the use of statistical measures which assume independent and essential normal distribution in an environment such as open, natural language where rare events are very common and distribution is anything but independent. Burrows (1992) and Stubbs (1995) point to the fact that not all word boundary collocations are possible, so tests such as the t-score must be giving excess weighting to observed word boundary collocations (since it is assumed that other collocations that are forbidden due to rules of syntax or by the domain of verbs and adjectives, are equally possible).

I intend to address these objections by using baseline data instead of calculated expected yields for my algorithm[17]. At present this is not possible as I cannot find a robust method to split out the lexicon into words that are comparable in terms of distribution. Two key aspects to consider are the frequency of each term in the pair, and also the nature of the distribution (see the discussion below on Church and Gale’s Inverse Document Frequency).

6.2 Reliance on Noun-Noun Pairs

I could not help noticing that I tend to choose noun-noun pairs when I test my software, and that the majority of pairs in papers on cooccurrence scoring that I have read are noun-noun pairs. I have chosen to formalise this tendency in this paper by restricting the terms to noun-noun pairs. It seems intuitive that certain types of words should fit better with the distributional hypothesis than others. Take the following extract for example:

Drug Safety
Yellow fever
Announcing plans to boost drug safety by finally allowing patients to report adverse reactions to drugs and vaccines, health chiefs said how disappointed they were at the public's apparent lack of interest in a pilot reporting scheme that ran in south-east London[18].

It is clear that some of the words used in this paragraph correlate with each other (and with the subject of the article given in the title), these include boost, drug, safety, patients, adverse reactions, vaccines, health chiefs, pilot. Others have no relation to the subject in hand (for example to, by, and, that, finally, disappointed). This suggests that some words (perhaps nouns in particular) will be suited to the application of the distributional hypothesis across wider windows, while others (especially function words) will not. This seems to make some intuitive sense, but it would be usefully if there were a way to prove it, and also a means of identifying such words other than simply sticking to nouns.

To this end I considered Inverse Document Frequency as described in Church and Gale (1999)[19]. Although the results in the paper look very promising my results with the BNC were inconclusive. I replicated Church and Gale’s experiment with 79 lemmas from the BNC with 1000 < f < 1020 which I ranked for IDF. The ten entries with the lowest IDF certainly looked to be ‘interesting’ (colitis, Claudia, Ellen, voltage, adjective, acid, Newman, MacDonald, rainbow, Kenya) whereas the ten with the highest IDF looked ‘uninteresting’ (invoke, no matter how, flick, reluctance, ample, flush, level,

[17] I used baseline data for the expected values in my earlier paper on cooccurrence scoring (Hardcastle 2001)

[18] Private Eye 1107, page 27

[19] IDF = $-\log_2 df_w/D$ where df_w is the document frequency of word w, and D is the total number of documents in the corpus. Note that the actual frequency of the word is not part of the calculation

flourish, stumble, obscure). However, the majority of the lemmas seemed to fall into a sizeable grey area in the middle.

Interestingly proper nouns and nouns seemed to be more interesting than adjectives and verbs, as below:

Tag	Average Frequency	Average IDF
NP0	1011	492
NN*	1010	573
AJ*	1009	672
VV*	1009	830

Table 7: IDF for terms of comparable frequency by class.

This is born out by listing the entries where the ratio of frequency over document frequency is highest. A list of the top 10,000 such entries is dominated by Proper Nouns and nouns, with some technical adjectives (e.g. nomic, methile, m-derived) and very few verbs. I tried to generalise the approach by finding some function of word frequency and document frequency (such as the one described above) that would allow me to compare words of different frequency but I have not yet been able to identify such a function.

So, although IDF produces some very interesting results I cannot identify a threshold, nor a test that would identify a threshold, to allow me to determine if a given word could sensibly be scored for word association using the distributional hypothesis.

6.3 Polysemy

Polysemy presents some problems which I have not explored in this paper. Take for example the pair pen-sheep or the pair pen-ink. Even if I restrict the search to occurrences of the noun pen, there are no instances of pen the noun which I would want to count for the first pairing and also for the second pairing. This means that both searches are counting cooccurrences (or a lack of them) based not just on x versus y but also on z versus y, where z is a homograph of x.

I have explored methods to disambiguate the word senses in the BNC elsewhere, but this is a very large and complex task. For now I think that it is enough to note that any word association scoring system based on measuring cooccurrences in a corpus will perform less well than it should unless word senses are disambiguated.

7 Appendices

7.1 Test 1 Source Data

Sets used for match and mismatch lists to determine the threshold for each measure.

Set A - types of clothes
jacket trouser jeans tie suit shirt dress sock hat shoe pants jumper blouse bra boot skirt
Set B - things relating to clothes
catwalk discount drier fashion foot iron launderette laundry leg magazine model office outdoor photo photograph rain sale school season shop shopping size store summer sun uniform waist winter writer
Set C - types of weather
sun rain snow spell shower cloud sky sleet hail mist fog weather ice
Set D - things relating to weather
north east south west england scotland wales hill mountain valley lake sea coast shore headland promontory motorway road traffic tv presenter chart map warning report alert news
Set E - maths, mathematics and mathematician
mathematics maths mathematician
Set F - things relating to maths
statistics equation formula calculator algebra domain quadrilateral teacher differentiation GCSE A-level integration class pupil pencil ruler protractor graph solution curve infinity exponential hypothesis
Set G - consultant, consultancy
consultant consultancy
Set H - things relating to consultancy
solution business internet computer computing government sector public finance mission target redundancy executive package strategy management tool employee leverage money wage fee

Match List

AxB CxD ExF GxH

Mismatch List

AxH CxF ExD GxB

7.2 Test 2 Source Data

Four sets of twenty hand-picked pairs used to test how robustly the measures could cope with looser associations and also to check decisiveness by measuring the overlap between positives and false positives.

Set 1 (Word Boundary)	Average Association Ratio	t-score	Inverse Distance	Dice Paragraph	Russian Doll
[Threshold]	**9.00**	**2.58**	**2.45**	**0.013**	**0.40**
tea-cup	9.51	160.23	4.59	0.205	0.98
tea-bag	9.05	75.94	4.89	0.014	0.72
coffee-cup	9.12	144.63	4.40	0.106	0.93
wine-bottle	10.90	127.03	4.14	0.123	1.00
bread-bin	8.37	31.84	3.93	0.006	0.81
traffic-jam	9.61	415.48	5.13	0.044	1.00
chocolate-bar	9.20	134.25	4.96	0.034	1.00
dessert-spoon	12.14	114.96	3.00	0.020	1.00
plant-pot	10.10	113.90	4.04	0.023	0.91
bus-lane	9.43	36.02	4.35	0.012	0.72
diamond-necklace	12.59	55.06	2.79	0.010	1.00
engagement-ring	8.94	141.10	3.78	0.021	1.00
cowboy-hat	9.40	93.59	3.24	0.011	1.00
exercise-book	7.60	19.49	4.60	0.009	0.36
shoulder-blade	9.61	174.75	4.91	0.022	1.00
apron-string	8.04	62.58	4.08	0.008	1.00
bath-mat	9.53	24.10	3.73	0.005	0.90
mouse-trap	8.60	29.02	2.98	0.006	0.79
mouse-mat	7.06	23.36	2.71	0.003	0.79
hub-cap	7.70	29.53	3.56	0.003	0.96
Average	**9.33**	**100.34**	**3.99**	**0.034**	**0.89**

Set 2 (Short Phrase)	Average Association Ratio	t-score	Inverse Distance	Dice Paragraph	Russian Doll
[Threshold]	**9.00**	**2.58**	**2.45**	**0.013**	**0.40**
coffee-tea	9.85	65.76	4.05	0.085	0.89
knife-fork	10.77	110.90	3.65	0.106	1.00
knife-spoon	10.31	19.82	2.55	0.020	0.99
doctor-patient	9.97	2.97	3.65	0.056	0.60
table-chair	8.83	21.57	3.86	0.053	0.59
wine-beer	11.06	41.81	3.61	0.064	0.95
cat-dog	9.95	33.97	3.80	0.051	0.77
mouse-cat	10.14	20.69	3.18	0.034	0.82
mouse-keyboard	11.39	28.38	2.94	0.035	1.00
mouse-hamster	11.18	17.96	2.22	0.008	1.00
hammer-nail	10.37	31.26	2.81	0.028	1.00
sugar-spice	11.61	15.88	2.68	0.014	1.00
pencil-paper	9.68	18.14	3.39	0.015	0.70
cradle-grave	9.70	32.76	2.48	0.020	1.00
penguin-keeper	12.17	24.05	2.30	0.002	1.00
pen-ink	12.54	84.15	3.26	0.062	1.00
doctor-nurse	9.97	33.15	3.85	0.050	0.74
lion-tiger	11.91	23.24	2.88	0.026	1.00
horse-saddle	11.76	29.42	3.08	0.017	1.00
horse-cow	9.38	3.92	2.95	0.010	0.54
Average	**10.63**	**32.99**	**3.16**	**0.038**	**0.88**

Set 3 (Associated)	Average Association Ratio	t-score	Inverse Distance	Dice Paragraph	Russian Doll
[Threshold]	**9.00**	**2.58**	**2.45**	**0.013**	**0.40**
doctor-hospital	9.68	23.71	3.93	0.045	0.61
chef-restaurant	12.45	33.01	2.89	0.028	1.00
cook-kitchen	10.64	11.20	2.68	0.013	0.86
cook-restaurant	9.60	0.00	1.20	0.004	0.54
farmer-tractor	10.63	5.27	2.13	0.007	0.75
waiter-tray	9.40	7.44	1.78	0.016	0.85
dentist-tooth	12.22	7.73	2.11	0.017	1.00
surgeon-operation	9.48	2.83	2.59	0.013	0.64
cow-milk	12.26	107.53	3.78	0.098	1.00
horse-rein	11.89	13.51	2.66	0.013	1.00
dog-hamster	9.20	0.21	1.62	0.001	0.55
dog-rabbit	9.60	0.00	2.62	0.009	0.55
pizza-salad	11.33	5.09	2.04	0.010	0.95
pizza-mushroom	10.27	2.17	1.36	0.006	0.83
pizza-olive	10.36	0.00	0.00	0.002	0.65
hat-wedding	8.72	1.42	2.25	0.006	0.48
pen-sheep	8.31	16.60	2.85	0.008	0.64
pen-writer	8.55	1.08	2.41	0.007	0.44
cage-monkey	10.36	1.54	1.65	0.005	0.74
penguin-zoo	13.82	5.94	1.15	0.009	1.00
Average	**10.44**	**12.31**	**2.19**	**0.016**	**0.75**

Set 4 (Nonsense)	**Average Association Ratio**	**t-score**	**Inverse Distance**	**Dice Paragraph**	**Russian Doll**
[Threshold]	**9.00**	**2.58**	**2.45**	**0.013**	**0.40**
bus-bottle	7.41	0.00	1.38	0.001	0.17
tea-exercise	7.03	0.00	1.08	0.002	0.14
shoulder-coffee	8.34	0.00	1.26	0.001	0.23
horse-wine	7.23	0.00	1.00	0.001	0.13
bath-sky	7.56	0.00	0.30	0.001	0.17
sugar-lane	6.51	0.00	0.85	0.001	0.13
cook-pencil	7.06	0.00	0.00	0.000	0.09
bath-race	5.83	0.00	0.48	0.000	0.05
tooth-keeper	6.08	0.00	0.00	0.000	0.08
blade-coffee	6.83	0.00	0.00	0.000	0.07
pencil-cow	6.65	0.00	0.00	0.000	0.06
mushroom-tie	6.89	0.00	0.00	0.000	0.07
cat-keyboard	5.88	0.00	0.78	0.000	0.09
hammer-cradle	8.30	0.00	0.00	0.000	0.15
dessert-iron	7.56	0.00	0.00	0.000	0.10
apron-lion	6.98	0.00	0.00	0.000	0.05
chocolate-surgeon	5.76	0.00	0.00	0.000	0.06
cap-salami	6.79	0.00	0.00	0.000	0.04
doctor-anchovy	6.74	0.00	0.00	0.000	0.04
tiger-pizza	4.76	0.00	0.00	0.000	0.01
Average	**6.81**	**0.00**	**0.36**	**0.000**	**0.10**

References

Budanitsky, A. and G. Hirst (2001) Semantic Distance in WordNet: An experimental, application-oriented evaluation of five measures. *NAACL 2001 Workshop on WordNet and Other Lexical Resources.*

Burrows, J. (1992) Computers and the study of literature, in C.S. Butler (ed.) *Computers and Written Texts*, 167-204.

Church, K. and W. Gale (1999) Inverse Document Frequency (IDF): a measure of deviations from Poisson, in *Natural Language Processing Using Very Large Corpora***,** 283-295.

Church, K. and P. Hanks (1990) Word association norms, mutual information and lexicography. *Computational Linguistics 16(1)*, 22-29.

Dagan, I. (n.d.) Contextual Word Similarity.

Dagan, I., L. Lee, et al. (1999. Similarity-based models of co-occurrence probablilities. *Machine Learning 34(1)*, 43-69.

Dunning, T. (1993) Accurate methods for statistics of surprise and coincidence. *Computational Linguistics 19*, 61-74.

Gale, K. Church, et al. (1993) A method for disambiguating word senses in a large corpus. *Computers and the Humanities 26*, 415-439.

Hardcastle, D. (2001) Using the BNC to produce dialectic cryptic crossword clues, *Corpus Linguistics 2001*, Lancaster.

Harris, Z. (1968) *Mathematical Structures of Language* (New York: Wiley).

Hunston, S. (2002) *Corpora in Applied Linguistics* (Cambridge: Cambridge University Press).

Jarmasz, M. and S. Spackowicz (2003) Roget's Thesaurus and Semantic Similarity. *Conference on Recent Advances in Natural Language Processing (RANLP 2003)*, Borovets, Bulgaria.

Manning, C. and H. Schutze (1999). Foundations of Statistical Natural Language Processing (Massachussetts Institute of Technology).

Matsumoto, Y. (2003) Lexical Knowledge Acquisition, in R. Mitkov (ed.) The Oxford Hanbdbook of Computational Linguistics (Oxford: Oxford University Press), 395-413.

Miller, G. and W. Charles (1991) Contextual correlates of semantic similarity. *Language and Cognitive Processes 6(1)*, 1-28.

Rubenstein, H. and J. Goodenough (1965) Contextual correlates of synonymy. *Communications of the ACM 8(10)*, 627-633.

Smadja, F., K. McKeown, et al. (1996) Translating collocations for bilingual lexicons: a statistical approach. *Computational Linguistics 22(1)*, 1-38.

Stubbs, M. (1995) Collocations and semantic profiles: on the cause of the trouble with qualitative studies. *Functions of Language 2,* 1-33.