Word Association Norms, Mutual Information, and Lexicography

Kenneth Ward Church
Bell Laboratories
Murray Hill, N.J.

Patrick Hanks
Collins Publishers
Glasgow, Scotland

Abstract

The term *word association* is used in a very particular sense in the psycholinguistic literature. (Generally speaking, subjects respond quicker than normal to the word *nurse* if it follows a highly associated word such as *doctor*.) We will extend the term to provide the basis for a statistical description of a variety of interesting linguistic phenomena, ranging from semantic relations of the doctor/nurse type (content word/content word) to lexico-syntactic co-occurrence constraints between verbs and prepositions (content word/function word). This paper will propose an objective measure based on the information theoretic notion of mutual information, for estimating word association norms from computer readable corpora. (The standard method of obtaining word association norms, testing a few thousand subjects on a few hundred words, is both costly and unreliable.) The proposed measure, *the association ratio*, estimates word association norms directly from computer readable corpora, making it possible to estimate norms for tens of thousands of words.

1. Meaning and Association

It is common practice in linguistics to classify words not only on the basis of their meanings but also on the basis of their co-occurrence with other words. Running through the whole Firthian tradition, for example, is the theme that ''You shall know a word by the company it keeps'' (Firth, 1957).

> ''On the one hand, *bank* co-occurs with words and expression such as *money, notes, loan, account, investment, clerk, official, manager, robbery, vaults, working in a, its actions, First National, of England*, and so forth. On the other hand, we find *bank* co-occurring with *river, swim, boat, east* (and of course *West* and *South*, which have acquired special meanings of their own), *on top of the*, and *of the Rhine*.'' [Hanks (1987), p. 127]

The search for increasingly delicate word classes is not new. In lexicography, for example, it goes back at least to the ''verb patterns'' described in Hornby's *Advanced Learner's Dictionary* (first edition 1948). What is new is that facilities for the computational storage and analysis of large bodies of natural language have developed significantly in recent years, so that it is now becoming possible to test and apply informal assertions of this kind in a more rigorous way, and to see what company our words do keep.

2. Practical Applications

The proposed statistical description has a large number of potentially important applications, including: (a) constraining the language model both for speech recognition and optical character recognition (OCR), (b) providing disambiguation cues for parsing highly ambiguous syntactic structures such as noun compounds, conjunctions, and prepositional phrases, (c) retrieving texts from large databases (e.g., newspapers, patents), (d) enhancing the productivity of computational linguists in compiling lexicons of lexico-syntactic facts, and (e) enhancing the productivity of lexicographers in identifying normal and conventional usage.

Consider the optical character recognizer (OCR) application. Suppose that we have an OCR device such as [Kahan, Pavlidis, Baird (1987)], and it has assigned about equal probability to having recognized *farm* and *form*, where the context is either: (1) *federal* ___ *credit* or (2) *some* ___ *of*.

- *federal* $\begin{bmatrix} farm \\ form \end{bmatrix}$ *credit*

- *some* $\begin{bmatrix} farm \\ form \end{bmatrix}$ *of*

The proposed association measure can make use of the fact that *farm* is much more likely in the first context and *form* is much more likely in the second to resolve the ambiguity. Note that alternative disambiguation methods based on syntactic constraints such as part of speech are unlikely to help in this case since both *form* and *farm* are commonly used as nouns.

3. Word Association and Psycholinguistics

Word association norms are well known to be an important factor in psycholinguistic research, especially in the area of lexical retrieval. Generally speaking, subjects respond quicker than normal to the word *nurse* if it follows a highly associated word such as *doctor*.

> ''Some results and implications are summarized from reaction-time experiments in which subjects either (a) classified successive strings of letters as words and nonwords, or (b) pronounced the strings. Both types of response to words (e.g., BUTTER) were consistently faster when preceded by associated words (e.g., BREAD) rather than unassociated words (e.g, NURSE).'' [Meyer, Schvaneveldt and Ruddy (1975), p. 98]

Much of this psycholinguistic research is based on empirical estimates of word association norms such as [Palermo and Jenkins (1964)], perhaps the most influential study of its kind, though extremely small and somewhat dated. This study measured 200 words by asking a few thousand subjects to write down a word after each of the 200 words to be measured. Results are reported in tabular form, indicating which words were written down, and by how many subjects, factored by grade level and sex. The word *doctor*, for example, is reported on pp. 98-100, to be most often associated with *nurse*, followed by *sick, health, medicine, hospital, man, sickness, lawyer*, and about 70 more words.

4. An Information Theoretic Measure

We propose an alternative measure, *the association ratio*, for measuring word association norms, based on the information theoretic concept of *mutual information*.[1] The proposed measure is more objective and less costly than the subjective method employed in [Palermo and Jenkins (1964)]. The association ratio can be scaled up to provide robust estimates of word association norms for a large portion of the language. Using the association ratio measure, the five most associated words are (in order): *dentists, nurses, treating, treat*, and *hospitals*.

---

1. This statistic has also been used by the IBM speech group [personal communication (1982)] for constructing language models for applications in speech recognition.

What is ''mutual information''? According to [Fano (1961), p. 28], if two points (words), x and y, have probabilities $P(x)$ and $P(y)$, then their mutual information, $I(x,y)$, is defined to be

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

Informally, mutual information compares the probability of observing x and y *together* (the joint probability) with the probabilities of observing x and y *independently* (chance). If there is a genuine association between x and y, then the joint probability $P(x,y)$ will be much larger than chance $P(x)\ P(y)$, and consequently $I(x,y) >> 0$. If there is no interesting relationship between x and y, then $P(x,y) \approx P(x)\ P(y)$, and thus, $I(x,y) \approx 0$. If x and y are in complementary distribution, then $P(x,y)$ will be much less than $P(x)\ P(y)$, forcing $I(x,y) << 0$.

In our application, word probabilities, $P(x)$ and $P(y)$, are estimated by counting the number of observations of x and y in a corpus, $f(x)$ and $f(y)$, and normalizing by N, the size of the corpus. (Our examples use a number of different corpora with different sizes: 15 million words for the 1987 AP corpus, 36 million words for the 1988 AP corpus, and 8.6 million tokens for the tagged corpus.) Joint probabilities, $P(x,y)$, are estimated by counting the number of times that x is followed by y in a window of w words, $f_w(x,y)$, and normalizing by N.

The window size parameter allows us to look at different scales. Smaller window sizes will identify fixed expressions (idioms such as *bread and butter*) and other relations that hold over short ranges; larger window sizes will highlight semantic concepts and other relationships that hold over larger scales.

The following table may help show the contrast.[2] In fixed expressions, such as *bread and butter* and *drink and drive*, the words of interest are separated by a fixed number of words and there is very little variance. In the 1988 AP, it was found that the two words are always exactly two words apart whenever they are found near each other (within five words). That is, the mean separation is two, and the variance is zero. Compounds also have very fixed word order (little variance), but the average separation is closer to one word rather than two. In contrast, relations such as *man/woman* are less fixed as indicated by a larger variance in their separation. (The nearly zero value for the mean separation for *man/women* indicates that words appear about equally often in either order.) Lexical relations come in several varieties. There are some like *refraining from* which are fairly fixed, others such as *coming from* which may be separated by an argument, and still others like *keeping from* which are almost certain to be separated by an argument.

Mean and Variance of the Separation Between X and Y

Relation	Word x	Word y	Separation	
			mean	variance
fixed	*bread*	*butter*	2.00	0.00
	drink	*drive*	2.00	0.00
compound	*computer*	*scientist*	1.12	0.10
	United	*States*	0.98	0.14
semantic	*man*	*woman*	1.46	8.07
	man	*women*	-0.12	13.08

2. Smadja (to appear) discusses the separation between collocates in a very similar way.

lexical	*refraining*	*from*	1.11	0.20
	coming	*from*	0.83	2.89
	keeping	*from*	2.14	5.53

The ideal window size is different for each case. For the remainder of this paper, the window size, w, will be set to 5 words as a compromise; this setting is large enough to show some of the constraints between verbs and arguments, but not so large that it would wash out constraints that make use of strict adjacency.[3]

Since the association ratio becomes unstable when the counts are very small, we will not discuss word pairs with $f(x,y) \leq 5$. An improvement would make use of t-scores, and throw out pairs that were not significant. Unfortunately, this requires an estimate of the variance of $f(x,y)$, which goes beyond the scope of this paper. For the remainder of this paper, we will adopt the simple but arbitrary threshold, and ignore pairs with small counts.

Technically, the *association ratio* is different from *mutual information* in two respects. First, joint probabilities are supposed to be symmetric: $P(x,y) = P(y,x)$, and thus, mutual information is also symmetric: $I(x,y) = I(y,x)$. However, the association ratio is not symmetric, since $f(x,y)$ encodes linear precedence. (Recall that $f(x,y)$ denotes the number of times that word x appears *before* y in the window of w words, not the number of times the two words appear in either order.) Although we could fix this problem by redefining $f(x,y)$ to be symmetric (by averaging the matrix with its transpose), we have decided not to do so, since order information appears to be very interesting. Notice the asymmetry in the pairs below (computed from 44 million words of 1988 AP text), illustrating a wide variety of biases ranging from sexism to syntax.

Asymmetry in 1988 AP Corpus (N = 44 million)

x	y	f(x, y)	f(y, x)
doctors	*nurses*	99	10
man	*woman*	256	56
doctors	*lawyers*	29	19
bread	*butter*	15	1
save	*life*	129	11
save	*money*	187	11
save	*from*	176	18
supposed	*to*	1188	25

Secondly, one might expect $f(x,y) \leq f(x)$ and $f(x,y) \leq f(y)$, but the way we have been counting, this needn't be the case if x and y happen to appear several times in the window. For example, given the sentence, ''Library workers were prohibited from saving books from this heap of ruins,'' which appeared in an AP story on April 1, 1988, $f(\textit{prohibited}) = 1$ and $f(\textit{prohibited}, \textit{from}) = 2$. This problem can be fixed by dividing $f(x,y)$ by $w-1$ (which has the consequence of subtracting $\log_2(w-1) = 2$ from our association ratio scores). This adjustment has the additional benefit of assuring that $\sum f(x,y) = \sum f(x) = \sum f(y) = N$.

When $I(x,y)$ is large, the association ratio produces very credible results not unlike those reported in

3. This definition $f_w(x,y)$ uses a rectangular window. It might be interesting to consider alternatives (e.g., a triangular window or a decaying exponential) that would weight words less and less as they are separated by more and more words. Other windows are also possible. For example, Hindle (Church, Gale, Hanks, Hindle 1989) has used a syntactic parser to select words in certain constructions of interest.

[Palermo and Jenkins (1964)], as illustrated in the table below. In contrast, when $I(x,y) \approx 0$, the pairs are less interesting. (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. One can make this statement precise by calibrating the measure with subjective measures. Alternatively, one could make estimates of the variance and then make statements about confidence levels, e.g., with 95% confidence, $P(x,y) > P(x)\ P(y)$.)

Some Interesting Associations with "Doctor"
in the 1987 AP Corpus (N = 15 million)

I(x, y)	f(x, y)	f(x)	x	f(y)	y
11.3	12	111	*honorary*	621	*doctor*
11.3	8	1105	*doctors*	44	*dentists*
10.7	30	1105	*doctors*	241	*nurses*
9.4	8	1105	*doctors*	154	*treating*
9.0	6	275	*examined*	621	*doctor*
8.9	11	1105	*doctors*	317	*treat*
8.7	25	621	*doctor*	1407	*bills*
8.7	6	621	*doctor*	350	*visits*
8.6	19	1105	*doctors*	676	*hospitals*
8.4	6	241	*nurses*	1105	*doctors*
Some Un-interesting Associations with "Doctor"					
0.96	6	621	*doctor*	73785	*with*
0.95	41	284690	*a*	1105	*doctors*
0.93	12	84716	*is*	1105	*doctors*

If $I(x,y) << 0$, we would predict that x and y are in complementary distribution. However, we are rarely able to observe $I(x,y) << 0$ because our corpora are too small (and our measurement techniques are too crude). Suppose, for example, that both x and y appear about 10 times per million words of text. Then, $P(x) = P(y) = 10^{-5}$ and chance is $P(x)\ P(x) = 10^{-10}$. Thus, to say that $I(x,y)$ is much less than 0, we need to say that $P(x,y)$ is much less than 10^{-10}, a statement that is hard to make with much confidence given the size of presently available corpora. In fact, we cannot (easily) observe a probability less than $1/N \approx 10^{-7}$, and therefore, it is hard to know if $I(x,y)$ is much less than chance or not, unless chance is very large. (In fact, the pair *a...doctors* above, appears significantly less often than chance. But to justify this statement, we need to compensate for the window size (which shifts the score downward by 2.0, e.g. from 0.96 down to – 1.04) and we need to estimate the standard deviation, using a method such as [Good (1953)].)[4]

5. Lexico-Syntactic Regularities

Although the psycholinguistic literature documents the significance of noun/noun word associations such as doctor/nurse in considerable detail, relatively little is said about associations among verbs, function words, adjectives, and other non-nouns. In addition to identifying semantic relations of the doctor/nurse variety, we believe the association ratio can also be used to search for interesting lexico-syntactic relationships between verbs and typical arguments/adjuncts. The proposed association ratio can be viewed as a formalization of Sinclair's argument:

4. Althougth the Good-Turing Method [Good (1953)] is more than 35 years old, it is still heavily cited. For example, [Katz (1987)] uses the method in order to estimate trigram probabilities in the IBM speech recognizer. The Good-Turing Method is helpful for trigrams that have not been seen very often in the training corpus.

> ''How common are the phrasal verbs with *set*? *Set* is particularly rich in making combinations with words like *about, in, up, out, on, off*, and these words are themselves very common. How likely is *set off* to occur? Both are frequent words; [*set* occurs approximately 250 times in a million words and] *off* occurs approximately 556 times in a million words... [T]he question we are asking can be roughly rephrased as follows: how likely is *off* to occur immediately after *set*? ... This is 0.00025×0.00055 [$P(x)\ P(y)$], which gives us the tiny figure of 0.0000001375 ... The assumption behind this calculation is that the words are distributed at random in a text [at chance, in our terminology]. It is obvious to a linguist that this is not so, and a rough measure of how much *set* and *off* attract each other is to compare the probability with what actually happens... *Set off* occurs nearly 70 times in the 7.3 million word corpus [$P(x,y) = 70/(7.3 \times 10^6) >> P(x)\ P(y)$]. That is enough to show its main patterning and it suggests that in currently-held corpora there will be found sufficient evidence for the description of a substantial collection of phrases... [Sinclair (1987)c, pp. 151-152]

Using Sinclair's estimates $P(set) \approx 250 \times 10^{-6}$, $P(off) \approx 556 \times 10^{-6}$ and $P(set,off) \approx 70/(7.3 \times 10^6)$, we would estimate the mutual information to be $I(set;off) = \log_2 P(set,off)/(P(set)\ P(off)) \approx 6.1$. In the 1988 AP corpus ($N = 44{,}344{,}077$), we estimate $P(set) \approx 13{,}046/N$, $P(off) \approx 20{,}693/N$ and $P(set,off) \approx 463/N$. Given these estimates, we would compute the mutual information to be $I(set;off) \approx 6.2$.

In this example, at least, the values seem to be fairly comparable across corpora. In other examples, we will see some differences due to sampling. Sinclair's corpus is a fairly balanced sample of (mainly British) text; the AP corpus is an unbalanced sample of American journalese.

This association between *set* and *off* is relatively strong; the joint probability is more than $2^6 = 64$ times larger than chance. The other particles that Sinclair mentions have association ratios of:

Some Phrasal Verbs in 1988 AP Corpus (N = 44 million)

x	y	f(x)	f(y)	f(x, y)	I(x; y)
set	*up*	13,046	64,601	2713	7.3
set	*off*	13,046	20,693	463	6.2
set	*out*	13,046	47,956	301	4.4
set	*on*	13,046	258,170	162	1.1
set	*in*	13,046	739,932	795	1.8
set	*about*	13,046	82,319	16	-0.6

The first three, *set up, set off* and *set out*, are clearly associated; the last three are not so clear. As Sinclair suggests, the approach is well suited for identifying phrasal verbs, at least in certain cases.

6. Preprocessing with a Part of Speech Tagger

Phrasal verbs involving the preposition *to* raise an interesting problem because of the possible confusion with the infinitive marker *to*. We have found that if we first tag every word in the corpus with a part of speech using a method such as [Church (1988)], and then measure associations between tagged words, we can identify interesting contrasts between verbs associated with a following preposition *to*/*in* and verbs associated with a following infinitive marker *to*/*to*. (Part of speech notation is borrowed from [Francis and Kucera (1982)]; in = preposition; to = infinitive marker; vb = bare verb; vbg = verb + ing; vbd = verb + ed; vbz = verb + s; vbn = verb + en.) The association ratio identifies quite a number of verbs associated in an interesting way with *to*; restricting our attention to pairs with a score of 3.0 or more, there are 768 verbs associated with the preposition *to*/*in* and 551 verbs with the infinitive marker *to*/*to*. The ten verbs found to

be most associated before *to*/*in* are:

- *to*/*in*: alluding/vbg, adhere/vb, amounted/vbn, relating/vbg, amounting/vbg, revert/vb, reverted/vbn, resorting/vbg, relegated/vbn

- *to*/*to*: obligated/vbn, trying/vbg, compelled/vbn, enables/vbz, supposed/vbn, intends/vbz, vowing/vbg, tried/vbd, enabling/vbg, tends/vbz, tend/vb, intend/vb, tries/vbz

Thus, we see there is considerable leverage to be gained by preprocessing the corpus and manipulating the inventory of tokens.

7. Preprocessing with a Parser

Hindle (Church, Gale, Hanks, Hindle 1989) has found it helpful to preprocess the input with the Fidditch parser [Hindle (1983a,b)] in order to identify associations between verbs and arguments, and postulate semantic classes for nouns on this basis. Hindle's method is able to find some very interesting associations as the next two tables demonstrate.

What can you drink?

Verb	Object	Mutual Info	Joint Freq
drink/V	*martinis/O*	12.6	3
drink/V	*cup_water/O*	11.6	3
drink/V	*champagne/O*	10.9	3
drink/V	*beverage/O*	10.8	8
drink/V	*cup_coffee/O*	10.6	2
drink/V	*cognac/O*	10.6	2
drink/V	*beer/O*	9.9	29
drink/V	*cup/O*	9.7	6
drink/V	*coffee/O*	9.7	12
drink/V	*toast/O*	9.6	4
drink/V	*alcohol/O*	9.4	20
drink/V	*wine/O*	9.3	10
drink/V	*fluid/O*	9.0	5
drink/V	*liquor/O*	8.9	4
drink/V	*tea/O*	8.9	5
drink/V	*milk/O*	8.7	8
drink/V	*juice/O*	8.3	4
drink/V	*water/O*	7.2	43
drink/V	*quantity/O*	7.1	4

What can you do to a telephone?

Verb	Object	Mutual Info	Joint Freq
sit_by/V	*telephone/O*	11.78	7
disconnect/V	*telephone/O*	9.48	7
answer/V	*telephone/O*	8.80	98
hang_up/V	*telephone/O*	7.87	3
tap/V	*telephone/O*	7.69	15
pick_up/V	*telephone/O*	5.63	11

return/V	*telephone/O*	5.01	19
be_by/V	*telephone/O*	4.93	2
spot/V	*telephone/O*	4.43	2
repeat/V	*telephone/O*	4.39	3
place/V	*telephone/O*	4.23	7
receive/V	*telephone/O*	4.22	28
install/V	*telephone/O*	4.20	2
be_on/V	*telephone/O*	4.05	15
come_to/V	*telephone/O*	3.63	6
use/V	*telephone/O*	3.59	29
operate/V	*telephone/O*	3.16	4

After running his parser over the 1988 AP corpus (44 million words), Hindle found $N = 4{,}112{,}943$ subject/verb/object (SVO) triples. The mutual information between a verb and its object was computed from these 4 million triples by counting how often the verb and its object were found in the same triple and dividing by chance. Thus, for example, *disconnect/V* and *telephone/O* have a joint probability of 7/N. In this case, chance is $84/N \times 481/N$ because there are 84 SVO triples with the verb *disconnect* and 481 SVO triples with the object *telephone*. The mutual information is $\log_2 7N/(84 \times 481) = 9.48$. Similarly, the mutual information for *drink/V beer/O* is $9.9 = \log 229N/(660 \times 195)$. (*drink/V* and *beer/O* are found in 660 and 195 SVO triples, respectively; they are found together in 29 of these triples.)

This application of Hindle's parser illustrates a second example of preprocessing the input in order to highlight certain constraints of interest. For measuring syntactic constraints, it may be useful to include some part of speech information and to exclude much of the internal structure of noun phrases. For other purposes, it may be helpful to tag items and/or phrases with semantic labels such as *person*, *place*, *time*, *body-part*, *bad*, etc.

8. Applications in Lexicography

Large machine-readable corpora are only just now becoming available to lexicographers. Up to now, lexicographers have been reliant either on citations collected by human readers, which introduced an element of selectivity and so inevitably distortion (rare words and uses were collected but common uses of common words were not), or on small corpora of only a million words or so, which are reliably informative for only the most common uses of the few most frequent words of English. (A million-word corpus such as the Brown Corpus is reliable, roughly, for only some uses of only some of the forms of around 4000 dictionary entries. But standard dictionaries typically contain twenty times this number of entries.)

The computational tools available for studying machine-readable corpora are at present still rather primitive. There are *concordancing* programs (see Figure 1 at the end of this paper), which are basically KWIC (key word in context [Aho, Kernighan, and Weinberger (1988), p. 122]) indexes with additional features such as the ability to extend the context, sort leftwards as well as rightwards, and so on. There is very little interactive software. In a typical situation in the lexicography of the 1980s, a lexicographer is given the concordances for a word, marks up the printout with colored pens in order to identify the salient senses, and then writes syntactic descriptions and definitions.

Although this technology is a great improvement on using human readers to collect boxes of citation index cards (the method Murray used in constructing the Oxford English Dictionary a century ago), it works well if there are no more than a few dozen concordance lines for a word, and only two or three main sense divisions. In analyzing a complex word such as *take, save*, or *from*, the lexicographer is trying to pick out significant patterns and subtle distinctions that are buried in literally thousands of concordance lines: pages and pages of computer printout. The unaided human mind simply cannot discover all the significant patterns, let alone group them and rank in order of importance.

The AP 1987 concordance to *save* is many pages long; there are 666 lines for the base form alone, and many more for the inflected forms *saved, saves, saving*, and *savings*. In the discussion that follows, we shall, for the sake of simplicity, not analyze the inflected forms and we shall only look at the patterns to the right of *save*.

Words Often Co-Occurring to the right of ''save''

I(x, y)	f(x, y)	f(x)	x	f(y)	y
9.5	6	724	*save*	170	*forests*
9.4	6	724	*save*	180	*$1.2*
8.8	37	724	*save*	1697	*lives*
8.7	6	724	*save*	301	*enormous*
8.3	7	724	*save*	447	*annually*
7.7	20	724	*save*	2001	*jobs*
7.6	64	724	*save*	6776	*money*
7.2	36	724	*save*	4875	*life*
6.6	8	724	*save*	1668	*dollars*
6.4	7	724	*save*	1719	*costs*
6.4	6	724	*save*	1481	*thousands*
6.2	9	724	*save*	2590	*face*
5.7	6	724	*save*	2311	*son*
5.7	6	724	*save*	2387	*estimated*
5.5	7	724	*save*	3141	*your*
5.5	24	724	*save*	10880	*billion*
5.3	39	724	*save*	20846	*million*
5.2	8	724	*save*	4398	*us*
5.1	6	724	*save*	3513	*less*
5.0	7	724	*save*	4590	*own*
4.6	7	724	*save*	5798	*world*
4.6	7	724	*save*	6028	*my*
4.6	15	724	*save*	13010	*them*
4.5	8	724	*save*	7434	*country*
4.4	15	724	*save*	14296	*time*
4.4	64	724	*save*	61262	*from*
4.3	23	724	*save*	23258	*more*
4.2	25	724	*save*	27367	*their*
4.1	8	724	*save*	9249	*company*
4.1	6	724	*save*	7114	*month*

It is hard to know what is important in such a concordance and what is not. For example, although it is easy to see from the concordance selection in Figure 1 that the word ''to'' often comes before ''save'' and the word ''the'' often comes after ''save,'' it is hard to say from examination of a concordance alone whether either or both of these co-occurrences have any significance.

Two examples will be illustrate how the association ratio measure helps make the analysis both quicker and more accurate.

8.1 Example 1: ''save ... from''

The association ratios (above) show that association norms apply to function words as well as content words. For example, one of the words significantly associated with *save* is *from*. Many dictionaries, for example Merriam-Webster's Ninth, make no explicit mention of *from* in the entry for *save*, although British learners' dictionaries do make specific mention of *from* in connection with *save*. These learners'

dictionaries pay more attention to language structure and collocation than do American collegiate dictionaries, and lexicographers trained in the British tradition are often fairly skilled at spotting these generalizations. However, teasing out such facts, and distinguishing true intuitions from false intuitions takes a lot of time and hard work, and there is a high probability of inconsistencies and omissions.

Which other verbs typically associate with *from*, and where does *save* rank in such a list? The association ratio identified 1530 words that are associated with *from*; 911 of them were tagged as verbs. The first 100 verbs are:

> refrain/vb, gleaned/vbn, stems/vbz, stemmed/vbd, stemming/vbg, ranging/vbg, stemmed/vbn, ranged/vbn, derived/vbn, ranged/vbd, extort/vb, graduated/vbd, barred/vbn, benefiting/vbg, benefitted/vbn, benefited/vbn, excused/vbd, arising/vbg, range/vb, exempts/vbz, suffers/vbz, exempting/vbg, benefited/vbd, prevented/vbd (7.0), seeping/vbg, barred/vbd, prevents/vbz, suffering/vbg, excluded/vbn, marks/vbz, profiting/vbg, recovering/vbg, discharged/vbn, rebounding/vbg, vary/vb, exempted/vbn, separate/vb, banished/vbn, withdrawing/vbg, ferry/vb, prevented/vbn, profit/vb, bar/vb, excused/vbn, bars/vbz, benefit/vb, emerges/vbz, emerge/vb, varies/vbz, differ/vb, removed/vbn, exempt/vb, expelled/vbn, withdraw/vb, stem/vb, separated/vbn, judging/vbg, adapted/vbn, escaping/vbg, inherited/vbn, differed/vbd, emerged/vbd, withheld/vbd, leaked/vbn, strip/vb, resulting/vbg, discourage/vb, prevent/vb, withdrew/vbd, prohibits/vbz, borrowing/vbg, preventing/vbg, prohibit/vb, resulted/vbd (6.0), preclude/vb, divert/vb, distinguish/vb, pulled/vbn, fell/vbn, varied/vbn, emerging/vbg, suffer/vb, prohibiting/vbg, extract/vb, subtract/vb, recover/vb, paralyzed/vbn, stole/vbd, departing/vbg, escaped/vbn, prohibited/vbn, forbid/vb, evacuated/vbn, reap/vb, barring/vbg, removing/vbg, stolen/vbn, receives/vbz.

Save ... from is a good example for illustrating the advantages of the association ratio. *Save* is ranked 319th in this list, indicating that the association is modest, strong enough to be important (21 times more likely than chance), but not so strong that it would pop out at us in a concordance, or that it would be one of the first things to come to mind.

If the dictionary is going to list *save ... from*, then, for consistency's sake, it ought to consider listing all of the more important associations as well. Of the 27 bare verbs (tagged 'vb') in the list above, all but 7 are listed in the Cobuild dictionary as occurring with *from*. However, this dictionary does not note that *vary, ferry, strip, divert, forbid,* and *reap* occur with *from*. If the Cobuild lexicographers had had access to the proposed measure, they could possibly have obtained better coverage at less cost.

8.2 Example 2: Identifying Semantic Classes

Having established the relative importance of *save ... from*, and having noted that the two words are rarely adjacent, we would now like to speed up the labor-intensive task of categorizing the concordance lines. Ideally, we would like to develop a set of semi-automatic tools that would help a lexicographer produce something like Figure 2, which provides an annotated summary of the 65 concordance lines for *save ... from*.[5] The *save ... from* pattern occurs in about 10% of the 666 concordance lines for *save*.

Traditionally, semantic categories have been only vaguely recognized, and to date little effort has been

5. The last unclassified line, *...save shoppers anywhere from $50...* raises interesting problems. Syntactic ''chunking'' shows that, in spite of its co-occurrence of *from* with *save*, this line does not belong here. An intriguing exercise, given the lookup table we are trying to construct, is how to guard against false inferences such as that since *shoppers* is tagged [PERSON], *$50 to $500* must here count as either BAD or a LOCATION. Accidental coincidences of this kind do not have a significant effect on the measure, however, although they do serve as a reminder of the probabilistic nature of the findings.

devoted to a systematic classification of a large corpus. Lexicographers have tended to use concordances impressionistically; semantic theorist, AI-ers, and others have concentrated on a few interesting examples, e.g., *bachelor*, and have not given much thought to how the results might be scaled up.

With this concern in mind, it seems reasonable to ask how well these 65 lines for *save ... from* fit in with all other uses of *save*? A laborious concordance analysis was undertaken to answer this question. When it was nearing completion, we noticed that the tags that we were inventing to capture the generalizations could in most cases have been suggested by looking at the lexical items listed in the association ratio table for *save*. For example, we had failed to notice the significance of time adverbials in our analysis of *save*, and no dictionary records this. Yet it should be clear from the association ratio table above that *annually* and *month*[6] are commonly found with *save*. More detailed inspection shows that the time adverbials correlate interestingly with just one group of *save* objects, namely those tagged [MONEY]. The AP wire is full of discussions of *saving $1.2 billion per month*; computational lexicography should measure and record such patterns if they are general, even when traditional dictionaries do not.

As another example illustrating how the association ratio tables would have helped us analyze the *save* concordance lines, we found ourselves contemplating the semantic tag ENV(IRONMENT) in order to analyze lines such as:

the trend to	save the forests[ENV]
it's our turn to	save the lake[ENV],
joined a fight to	save their forests[ENV],
can we get busy to	save the planet[ENV]?

If we had looked at the association ratio tables before labeling the 65 lines for *save ... from*, we might have noticed the very large value for *save ... forests*, suggesting that there may be an important pattern here. In fact, this pattern probably subsumes most of the occurrences of the ''*save* [ANIMAL]'' pattern noticed in Figure 2. Thus, tables do not provide semantic tags, but they provide a powerful set of suggestions to the lexicographer for what needs to be accounted for in choosing a set of semantic tags.

It may be that everything said here about *save* and other words is true only of 1987 American journalese. Intuitively, however, many of the patterns discovered seem to be good candidates for conventions of general English. A future step would be to examine other more balanced corpora and test how well the patterns hold up.

9. Conclusions

We began this paper with the psycholinguistic notion of word association norm, and extended that concept toward the information theoretic definition of mutual information. This provided a precise statistical calculation that could be applied to a very large corpus of text in order to produce a table of associations for tens of thousands of words. We were then able to show that the table encoded a number of very interesting patterns ranging from *doctor ... nurse* to *save ... from*. We finally concluded by showing how the patterns in the association ratio table might help a lexicographer organize a concordance.

In point of fact, we actually developed these results in basically the reverse order. Concordance analysis is

6. The word *time* itself also occurs significantly in the table, but on closer examination it is clear that this use of *time* (e.g., *to save time*) counts as something like a commodity or resource, not as part of a time adjunct. Such are the pitfalls of lexicography (obvious when they are pointed out).

still extremely labor-intensive, and prone to errors of omission. The ways that concordances are sorted don't adequately support current lexicographic practice. Despite the fact that a concordance is indexed by a single word, often lexicographers actually use a second word such as *from* or an equally common semantic concept such as a time adverbial to decide how to categorize concordance lines. In other words, they use two words to *triangulate in* on a word sense. This triangulation approach clusters concordance lines together into word senses based primarily on usage (distributional evidence), as opposed to intuitive notions of meaning. Thus, the question of what is a word sense can be addressed with syntactic methods (symbol pushing), and need not address semantics (interpretation), even though the inventory of tags may appear to have semantic values.

The triangulation approach requires ''art.'' How does the lexicographer decide which potential cut points are ''interesting'' and which are merely due to chance? The proposed association ratio score provides a practical and objective measure which is often a fairly good approximation to the ''art.'' Since the proposed measure is objective, it can be applied in a systematic way over a large body of material, steadily improving consistency and productivity.

But on the other hand, the objective score can be misleading. The score takes only distributional evidence into account. For example, the measure favors *set ... for* over *set ... down*; it doesn't know that the former is less interesting because its semantics are compositional. In addition, the measure is extremely superficial; it cannot cluster words into appropriate syntactic classes without an explicit preprocess such as Church's parts program or Hindle's parser. Neither of these preprocesses, though, can help highlight the ''natural'' similarity between nouns such as *picture* and *photograph*. Although one might imagine a preprocess that would help in this particular case, there will probably always be a class of generalizations that are obvious to an intelligent lexicographer, but lie hopelessly beyond the objectivity of a computer.

Despite these problems, the association ratio could be an important tool to aid the lexicographer, rather like an index to the concordances. It can help us decide what to look for; it provides a quick summary of what company our words do keep.

References

Church, K., (1988), ''A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text,'' Second Conference on Applied Natural Language Processing, Austin, Texas.

Church, K, Gale, W., Hanks, P., Hindle, D., (1989) ''Parsing, Word Associations and Typical Predicate-Argument Relations,'' International Workshop on Parsing Technologies, CMU.

Fano, R., (1961), *Transmission of Information,* MIT Press, Cambridge, Massachusetts.

Firth, J., (1957), ''A Synopsis of Linguistic Theory 1930-1955'' in *Studies in Linguistic Analysis*, Philological Society, Oxford; reprinted in Palmer, F., (ed. 1968), *Selected Papers of J.R. Firth*, Longman, Harlow.

Francis, W., and Kucera, H., (1982), *Frequency Analysis of English Usage,* Houghton Mifflin Company, Boston.

Good, I. J., (1953), *The Population Frequencies of Species and the Estimation of Population Parameters,* Biometrika, Vol. 40, pp. 237-264.

Hanks, P. (1987), ''Definitions and Explanations,'' in Sinclair (1987b).

Hindle, D., (1983a), ''Deterministic Parsing of Syntactic Non-fluencies,'' ACL Proceedings.

Hindle, D., (1983b), ''User manual for Fidditch, a deterministic parser,'' Naval Research Laboratory

Hornby, A., (1948), *The Advanced Learner's Dictionary,* Oxford University Press.

Kahan, S., Pavlidis, T., and Baird, H., (1987) ''On the Recognition of Printed Characters of any Font or Size,'' IEEE Transactions PAMI, pp. 274-287.

Katz, S. M., (1987), ''Estimation of probabilities from sparse data for the language model component of a speech recognizer,'' *IEEE Transactions on Acoustics, Speech, and Signal Processing,* v. ASSP-35, pp. 400-401.

Meyer, D., Schvaneveldt, R., and Ruddy, M., (1975), ''Loci of Contextual Effects on Visual Word-Recognition,'' in Rabbitt, P., and Dornic, S., (eds.), *Attention and Performance V*, Academic Press, London, New York, San Francisco.

Palermo, D., and Jenkins, J., (1964) ''Word Association Norms,'' University of Minnesota Press, Minneapolis.

Sinclair, J., Hanks, P., Fox, G., Moon, R., Stock, P. (eds), (1987a), *Collins Cobuild English Language Dictionary,* Collins, London and Glasgow.

Sinclair, J., (1987b), ''The Nature of the Evidence,'' in Sinclair, J. (ed.), *Looking Up: an account of the COBUILD Project in lexical computing,* Collins, London and Glasgow.

Smadja, F., (to appear) ''Microcoding the Lexicon with Co-Occurrence Knowledge,'' in Zernik (ed), ''Lexical Acquisition: Using on-line Resources to Build a Lexicon,'' MIT Press.

Figure 1: Short Sample of the Concordance to ‘‘Save’’ from the AP 1987 Corpus

rs Sunday, calling for greater economic reforms to	save China from poverty.
mmission asserted that ‘‘ the Postal Service could	save enormous sums of money in contracting out individual c
Then, she said, the family hopes to	save enough for a down payment on a home.
e out-of-work steelworker, ‘‘ because that doesn’t	save jobs, that costs jobs. ’’
‘‘ We suspend reality when we say we’ll	save money by spending $10,000 in wages for a public works
scientists has won the first round in an effort to	save one of Egypt’s great treasures, the decaying tomb of R
about three children in a mining town who plot to	save the ‘‘ pit ponies ’’ doomed to be slaughtered.
GM executives say the shutdowns will	save the automaker $500 million a year in operating costs a
rtment as receiver, instructed officials to try to	save the company rather than liquidate it and then declared
The package, which is to	save the country nearly $2 billion, also includes a program
newly enhanced image as the moderate who moved to	save the country.
million offer from chairman Victor Posner to help	save the financially troubled company, but said Posner stil
after telling a delivery-room doctor not to try to	save the infant by inserting a tube in its throat to help i
h birthday Tuesday, cheered by those who fought to	save the majestic Beaux Arts architectural masterpiece.
at he had formed an alliance with Moslem rebels to	save the nation from communism.
‘‘ Basically we could	save the operating costs of the Pershings and ground-launch
We worked for a year to	save the site at enormous expense to us, ’’ said Leveillee.
their expensive mirrors, just like in wartime, to	save them from drunken Yankee brawlers, ’’ Tass said.
ard of many who risked their own lives in order to	save those who were passengers. ’’
We must increase the amount Americans	save. ’’

Figure 2: Some AP 1987 Concordance lines to ‘save ... from,’ roughly sorted into categories

save X from Y (65 concordance lines)

1 save PERSON from Y (23 concordance lines)

1.1 save PERSON from BAD (19 concordance lines)

(Robert DeNiro) to	save Indian tribes[PERSON] from genocide[DESTRUCT[BAD]] at the hands of
‘‘ We wanted to	save him[PERSON] from undue trouble[BAD] and loss[BAD] of money , ’’
Murphy was sacrificed to	save more powerful Democrats[PERSON] from harm[BAD] .
‘‘ God sent this man to	save my five children[PERSON] from being burned to death[DESTRUCT[BAD]] and
Pope John Paul II to ‘‘	save us[PERSON] from sin[BAD] . ’’

1.2 save PERSON from (BAD) LOC(ATION) (4 concordance lines)

rescuers who helped	save the toddler[PERSON] from an abandoned well[LOC] will be feted with a parade
while attempting to	save two drowning boys[PERSON] from a turbulent[BAD] creek[LOC] in Ohio[LOC]

2. save INST(ITUTION) from (ECON) BAD (27 concordance lines)

member states to help	save the EEC[INST] from possible bankruptcy[ECON][BAD] this year .
should be sought ’’ to	save the company[CORP[INST]] from bankruptcy[ECON][BAD] .
law was necessary to	save the country[NATION[INST]] from disaster[BAD] .
operation ’’ to	save the nation[NATION[INST]] from Communism[BAD][POLITICAL] .
were not needed to	save the system from bankruptcy[ECON][BAD] .
his efforts to	save the world[INST] from the likes of Lothar and the Spider Woman

3. save ANIMAL from DESTRUCT(ION) (5 concordance lines)

give them the money to	save the dogs[ANIMAL] from being destroyed[DESTRUCT] ,
program intended to	save the giant birds[ANIMAL] from extinction[DESTRUCT] ,

UNCLASSIFIED (10 concordance lines)

walnut and ash trees to	save them from the axes and saws of a logging company .
after the attack to	save the ship from a terrible[BAD] fire , Navy reports concluded Thursday .
certificates that would	save shoppers[PERSON] anywhere from $50[MONEY] [NUMBER] to $500[MONEY] [NUMBER]