# Unsupervised learning of multi-word verbs

Don Blaheta and Mark Johnson
{dpb,mj}@cs.brown.edu
Brown University

## Abstract

Collocation is a linguistic phenomenon that is difficult to define and harder to explain; it has been largely overlooked in the field of computational linguistics due to its difficulty. Although standard techniques exist for finding collocations, they tend to be rather noisy and suffer from sparse data problems. In this paper, we demonstrate that by utilising parsed input to concentrate on one very specific type of collocation—in this case, verbs with particles, a subset of the so-called "multi-word" verbs—and applying an algorithm to promote those collocations in which we have more confidence, the problems with statistically learning collocations can be overcome.

## 1 Introduction

### 1.1 Collocations

A collocation, in the most general sense, is just some number of words that tend to occur together often; a native speaker would probably say that they fit together well. There is some disagreement in the literature as to what *exactly* should be defined as a collocation—in particular, whether totally opaque constructions count. Manning and Schütze (1999) explicitly consider idioms to be a kind of on collocation; Cruse (1986) sets up a contrast between the semantically opaque idiom and the semantically transparent collocation.

In any case, it is agreed that the two main necessary qualities of collocations are as follows:

- They are not fully compositional. At the very least, they carry extra connotation, or one or more of the constituent words has a restricted or modified definition when within the collocation.
- They are not easily modifiable. While many collocations can have synonyms substituted in, or modifiers added in the middle, and still "make sense", native speakers will find such a construction to be understandable but odd.

Multi-word verbs comprise a domain that definitely meets both these criteria.

### 1.2 Multi-word verbs

For this paper, we will essentially use the definition of multi-word verbs given in (Quirk et al., 1985). Quirk defines a multi-word verb (MWV) as "a unit which behaves to some extent either lexically or syntactically as a single verb", as distinct from those MWV-like constructions that are freely variable (i.e. fully compositional). They are comprised of one verb and one or more other words, which may be of any class: "rely on", "take care of", "see fit". This work restricts itself to MWVs comprised of a verb and some number of *particles*, which may be either prepositions or adverbs.[1]

### 1.3 Idiomaticity vs. Compositionality

Cruse (1986) defines an idiom as "an expression whose meaning cannot be inferred from the meanings of its parts". He then goes on to give two necessary properties of idioms: that they be lexically complex, and that they comprise indivisible semantic units. Multi-word verbs span the continuum between the nearly transparent compositional phrase and the fully opaque idiom. It is often difficult to classify them exactly, due to the small number of particles available to contrast with, and the even smaller number of particles that fit the pragmatics of a given situation. Nonetheless, it should be fairly clear that some MWVs can be further broken down into semantic units, as demonstrated by synonym and related-word substitutability:

$$\text{agree on} \quad \text{vs.} \quad \left.\begin{array}{r}\text{agree}\\ \text{disagree}\\ \text{expound}\\ \text{lecture}\end{array}\right\}\left\{\begin{array}{l}\text{on}\\ \text{about}\\ \text{regarding}\\ \text{concerning}\end{array}\right. \tag{1}$$

[1] Note that this definition subsumes that used in the Treebank literature (Bies et al., 1995), wherein the word 'particle' refers only to the standalone adverbs, such as 'out' and 'together'.

while for others the obvious synonyms and related particles don't work as well:

$$\text{rely on} \quad \text{vs.} \quad \text{*rely}\left\{\begin{array}{l}\text{over}\\ \text{in}\\ \text{about}\end{array}\right. \tag{2}$$

However, even these MWVs can exhibit some amount of divisibility within highly constrained environments:

$$\text{rely on} \quad \text{vs.} \quad \left.\begin{array}{c}\text{rely}\\ \text{count}\\ \text{depend}\\ \text{hinge}\end{array}\right\}\left\{\begin{array}{l}\text{on}\\ \text{upon}\end{array}\right. \tag{3}$$

There are two available views of this sort of highly constrained alternation: on the one hand, we could argue that the particles participating in these MWVs (in the case of example (3), 'on' and 'upon') have many distinct senses, each of which (or at least most of which) strongly select for a few specific verbs (in this case, 'rely', 'count', etc.). On the other hand, we could also argue that this seemingly semi-productive alternation is pure coincidence—or at best historically related—and that synchronically these constructions are idiomatic. The resolution of this point is not within the scope of this paper.

## 2 Method

### 2.1 Input and preprocessing

Any statistical work on collocation clearly needs to have copious amounts of input in order to be of any use. In the past, that has forced the use of raw text input—at best, tagged input—as there did not exist any more fully annotated corpora of sufficient size. However, 30 million words of parsed Wall Street Journal text make the BLLIP'99 corpus (Charniak et al., 1999) large enough to make lexical information gathering a somewhat reasonable endeavour.

From this parsed data, we tallied every verb-particle frame: for every verb in the corpus (any word whose part-of-speech tag started with VB), we recorded it along with the heads of any siblings labelled PRT ("particle" in the Penn Treebank sense; what this paper refers to as an "adverbial particle") or PP ("prepositional phrase"). A given instance of a verb could have zero particle siblings, or in extreme cases as many as five or six. To further alleviate the ever-present sparse data problem, we then stemmed the verbs using a relatively naïve algorithm (Porter, 1980), augmented with knowledge of about eighty common irregular verb forms.

### 2.2 Mutual information

One of the more popular statistical methods for ranking collocations is *mutual information*, as described by Church and Hanks (1989) among others. However, like most of the other lexical statistical methods, it is quite sensitive to sparse data problems, tending to promote low-frequency items. One attempt at combatting this tendency was given by Dunning (1993); more recently, Johnson (2001) has proposed a "confidence interval estimator" that is fairly successful.

There are, however, at least two problems with it in this domain. First of all, it doesn't take into account multi-word verbs of length greater than two. At best, it could count such constructions as a separate 2-word verb for each involved particle—"live up" and "live to" for "live up to"—but then each dilutes the relative probability of the other.

The second problem is how exactly to account for verbs that occur without any particle siblings. If we do not count them at all, the most prominent verb-particle pairs tend to include verbs that occur with or without a particle (but when they do appear with a particle, it is generally the same one). The clearest collocations, however, generally are those verbs which hardly ever occur without a given particle; thus, ideally, we should factor in the "no particle" occurrences of verbs.

We have developed a generalisation of the work in Johnson (2001) to generate confidence intervals for $n$-grams. This algorithm, like the old one, discounts the probability of low-frequency items; in addition, though, it estimates the likelihood of seeing given $n$-grams while discounting those $n$-grams that are only likely due to their component parts.

### 2.3 Log-linear models of $n$-way interaction

This section describes the statistics we used as measures of association to find the strongly associated verb-particle tuples. In fact, the techniques are general, in that they provide a measure of how strongly an $n$-tuple of binary variables are associated, and it is not important that these variables represent the occurrences of particular words.

We propose two different measures of association $\mu$ and $\mu_1$, which we call the "all subtuples" and "unigram subtuples" measures below. As we explain below, they seem to identify very different kinds of collocations, so both are useful in certain circumstances. These measures are estimates of $\lambda$ and $\lambda_1$ respectively, which are particular parameters of certain log-linear models. In cases where the counts are small the estimates of $\lambda$ and $\lambda_1$

may be noisy, and so high values from small count data should be discounted in some way when being compared with values from large count data. We do this by also estimating the asymptotic standard error $\sigma$ and $\sigma_1$ of $\lambda$ and $\lambda_1$ respectively, and set $\mu = \lambda - 3.29\sigma$ and $\mu_1 = \lambda_1 - 3.29\sigma_1$. This corresponds to setting the measures $\mu$ and $\mu_1$ to the lower bound of a 0.001 confidence interval for $\lambda$ and $\lambda_1$ respectively, which is a systematic way of trading recall for precision in the face of noisy data (Johnson, 2001).

Now we turn to the estimation of $\lambda, \lambda_1, \sigma$ and $\sigma_1$. Let $X_1, \ldots, X_n$ be random variables, where each $X_i$ ranges over the $i$ component of the tuples. In our application, $X_1$ ranges over verbs and each $X_i$, $2 \leq i \leq n$, ranges over particles and a distinguished null symbol '$\Box$' which is used to "fill" the value of $X_i, i > j+1$ when the verb appears with only $j$ following prepositions or particles. For example, if $n = 3$, then the verb phrase *write it off* would correspond to the tuple $X_1 = \textit{write}, X_2 = \textit{off}, X_3 = \Box$.

Suppose we wish to measure the association of the tuple $X_1 = x_1 \wedge \ldots \wedge X_n = x_n$. The measures we propose are defined in terms of the number of times the possible conjunctions of equalities and inequalities of variables occur in the training data. We represent each possible combination of these equalities and inequalities with an $n$-bit integer $b, 0 \leq b < 2^n$ as follows: the conjunction of equalities and inequalities represented by $b$ contains the equality $X_i = x_i$ iff the $i$th bit of $b$ is 1, and it contains the inequality $X_i \neq x_i$ iff the $i$th bit of $b$ is 0. Thus $b = 0$ represents the conjunction $X_1 \neq x_1 \wedge \ldots X_n \neq x_n$ and $b = 2^n - 1$ represents $X_1 = x_1 \wedge \ldots \wedge X_n = x_n$. Continuing with the example above, $b = 3$ represents the conjunction of equalities and inequalities $X_1 = \textit{write} \wedge X_2 = \textit{off} \wedge X_3 \neq \Box$.

For a fixed tuple of values $x_1, \ldots, x_n$ let $c_b$ be the number of times the conjunction of equalities and inequalities represented by $b$ is true in the training data. Continuing with the example, $c_3$ would be the number of times *write* was observed followed by *off* and some other preposition or particle. (In fact, following the suggestion in Goodman (1970), we add $\frac{1}{2}$ to each $c_b$ as a continuity correction for small counts; this also avoids overflow problems with zero counts). Further, let $\#(b)$ be the number of bits set to 1 in $b$. Then the quantities $\lambda, \lambda_1, \sigma$ and $\sigma_1$ are given by:

$$\lambda = \sum_{b=0}^{2^n-1} (-1)^{n-\#(b)} \log c_b$$

$$\lambda_1 = \log c_{2^n-1} - \sum_{\#(b)=1} \log c_b + (n-1)\log c_0$$

$$\sigma = \sqrt{\sum_{b=0}^{2^n-1} \frac{1}{c_b}}$$

$$\sigma_1 = \sqrt{\frac{1}{c_{2^n-1}} + \sum_{\#(b)=1} \frac{1}{c_b} + \frac{(n-1)^2}{c_0}}$$

These formulae for $\lambda$ and $\lambda_1$ are maximum likelihood estimators for the $n$-way interaction terms in certain saturated log-linear models. Log-linear models provide a general framework for constructing models under various assumptions about which combinations of variables interact, in this sense they are like the more well-known ANOVA models. Unlike ANOVA models, log-linear models do not assume that the data is normally distributed; rather, log-linear models fit a multinomial or Poisson distribution, which should result in a better fit to count data (Agresti, 1996).

Both $\lambda$ and $\lambda_1$ are the $n$-way interaction term in saturated log-linear models for the count data $c$. This term is in effect the difference between the log of the count $c_{2^n-1}$ and the log count that would be expected given the lower order interactions in the model.

The parameter $\lambda$ is the $n$-way interaction term in a log-linear model which also contains all lower-order (i.e., $0, 1, \ldots, n-1$ way) interaction terms; thus we call $\mu$ (the lower bound of the confidence interval for $\lambda$) the *all subtuples measure*. The parameter $\lambda_1$ is the $n$-way interaction term in a log-linear model which also contains only 0 and 1-way interaction terms; hence we call $\mu_1$ (the lower bound on the confidence interval for $\lambda_1$) the *unigram subtuples measure*.

For the special case of $n = 2$, the two measures are identical: $\lambda = \lambda_1$ and $\sigma = \sigma_1$. In fact, $\lambda$ is just the log odds ratio, and $\sigma$ is is its asymptotic standard error (Hollander and Wolfe, 1999). Thus both of these statistics can be regarded as different generalisations of the odds ratio for $n$-way interactions.

For $n > 2$ the two measures can behave quite differently. Although this paper is not directly concerned with general word $n$-tuples, the difference between these two measures is perhaps clearest with them. The word-tuples which score highest on the $\lambda_1$ measure are typically multi-word names, such as *Drexel Burnham Lambert* and *Ho Chi Minh*. In our training corpus, the words in such names only occur in these particular names; hence the tuple probability much larger than the product of the unigram word probabilities, and

| | |
|---|---|
| consist of | |
| fend off | |
| pale beside | |
| ward off | |
| accord to | *from "according to"* |
| cobble together | |
| shy away | |
| revolve around | |
| fritter away | |
| dispose of | |
| bog down | |
| swallow whole | |
| accuse of | |
| beef up | |
| spun off | *reported distinct from "spin* |
| latch onto | *off" due to faulty stemmer* |
| bail out | |
| yield less | |
| single out | |
| scale back | |
| lag behind | |
| squirrel away | |
| perch atop | |
| stave off | |
| shore up | |

Figure 1: Top 25 2-word verbs

so the tuple receives a high $\lambda_1$ score. On the other hand, the tuple probability is completely predictable from the word bigram probabilities (e.g., given $X_1 = \textit{Drexel}$ it is completely predictable that $X_2 = \textit{Burnham}$ and $X_3 = \textit{Lambert}$), so name-like word-triples typically score low on the $\lambda$ measure. (Word triples such as *little or no*, *by and large*, *let go of*, etc., score highly on the $\lambda$ measure).

We now briefly sketch the origins of the formulae above. The formulae for $\lambda$ and $\sigma$ are from Goodman (1970). We derived the formulae for $\lambda_1$ and $\sigma_1$ from the maximum likelihood equations for log linear models presented in Agresti (1990). Because we fit saturated models, the estimated parameters are always linear combinations of the log count data. On the other hand, the reader will notice that the parameter estimates are *differences* of log counts, and so become progressively more sensitive to noise in the count data as $n$ increases. In practice, we find that the number of tuples for which the lower bound of the confidence interval is greater than some positive constant drops quickly as $n$ increases, and quality of the tuples retrieved also decreases.

## 3 Results

The top 25 2-word verbs are reported in Figure 1. They have been un-stemmed by hand for readability (actual output included "cobbl together", "shi away", and so on), and annotated in italics where necessary. It is worth noting that although we only report the top 25 here, the collocations are almost all "good" for several hundred, and good collocations continue to appear with great frequency well into the thousands.

For three-word verbs, there were two different sets of output: one using the all subtuples measure and one using the unigram subtuples measure. The top 25 MWVs from each are reported here, in Figures 3 and 4, respectively. The three-word verbs are not quite as prolific as the two-word verbs, of course, but they do continue well beyond the 25 we give here.

## 4 Evaluation

We evaluated our results according to the judgements of native speakers of English regarding the relatedness of each $n$-gram. To get a sense of how the quality of the output degrades, we looked not just at the top hundred, but also at the groups of a hundred that started after one thousand and two thousand tuples. To minimise the bias of the judges, we combined all the $n$-grams they were to evaluate into a single file, randomised its order, and then parcelled out sections for each to consider.

We asked our evaluators to judge each item on four criteria (see Figure 2). Three—phrasality, transitivity, and opacity—are (relatively) objective, while the fourth—relatedness—is completely subjective. The objective judgements are used primarily to indicate just what sort of collocation our algorithm recovers, reported in percentages of the data group falling in a given category, while the subjective judgement is used as our reportable success rate, reported as an average of all "goodness" ratings in the evaluation group..

Many of the tuples were evaluated by multiple judges. If their judgements differed, they were averaged together; Table 4 gives some statistics on interannotator agreement. The agreement metric is the percentage of all doubly-evaluated tuples on which the judges agreed; for the relatedness judgement, an exact-agreement percentage is given, but the average difference between judgements should also be considered.

It is probably worth noting at this point that the evaluators were all college-educated native speakers of English, but with varying degrees of linguistic training. As such, their relatedness judgements should be fairly trustworthy, but the objective MWV typology may be slightly suspect; each was given a set of guidelines and explanations, but the agreement statistic indicates that there was considerable disagreement even over the

| Ranks | Phrasal? | Transitive? | Opaque? | Good collocation? |
|---|---|---|---|---|
| 1–100 | 46% | 66% | 49% | 3.84 |
| 1001–1100 | 29% | 55% | 32% | 3.11 |
| 2001–2100 | 29% | 61% | 30% | 2.84 |

Table 1: Evaluation of bigram output

**Phrasality.** This is the phrasal/prepositional distinction found in (Quirk et al., 1985): is the particle an adverb (phrasal) or a preposition? In 3-word verbs, the two particles are judged separately.

**Transitivity.** A multi-word verb is transitive if it has a direct object (also as defined in (Quirk et al., 1985); we do not include prepositional objects in this count).

**Opacity.** A tuple is considered "opaque" if its meaning cannot be guessed from the meanings of its parts.

**Relatedness.** A purely subjective judgement on a scale from 1–5, on whether a collocation really is strongly related or not.

Annotators were allowed to skip a tuple if they could not think of how such a collocation would occur.

Figure 2: Judgement criteria

leave over from
rang in from
make up of
accuse by of
face up to
trade among for
miss out on
think of as
sign off on
receive by at
follow through on
bar by from
ask on to
hold on to
reach out to
sit across from
said because in
end with to
include from from
total off from
urge on by
own up to
file for from
look forward to
help along by

Figure 3: Top 25 3-word verbs, all subtuples

bail out of
spill over into
line up behind
shy away from
spin off into
spun off into *duplicate due to faulty stemmer*
push ahead with
parcel out among
divvy up among
bog down amid
single out as
consist of of
bog down over
sprung up around
clamp down on
single out for
bog down in
branch out into
fend off by
redeem at plus
crack down on
bog down by
spin off to
square off against
team up with

Figure 4: Top 25 3-word verbs, unigram subtuples

objective classifications.

## 5 Future Work

Although we have successfully collected a large list of multi-word verbs, it has no particular ordering or subdivision. One possible extension to this work would be to compute not just the words themselves but also the specific type of MWV they comprise—transitive or intransitive, phrasal or prepositional. This is nontrivial, as intervening noun phrases can be either direct objects or adverbial modifiers (like "this evening"), and the main test for phrasality is done by performing a transformation on the sentence and using native speaker judgement as to whether the result is meaningful.

Another area of work is in dealing more explicitly with passive constructions. In the current

| Ranks | Phrasal? | Transitive? | Opaque? | Good collocation? |
|---|---|---|---|---|
| 1–100 | 41%/10% | 61% | 44% | 2.57 |
| 1001–1100 | 26%/4% | 74% | 24% | 2.12 |

Table 2: Evaluation of trigram output: all subtuples

| Ranks | Phrasal? | Transitive? | Opaque? | Good collocation? |
|---|---|---|---|---|
| 1–100 | 65%/4% | 62% | 73% | 3.76 |
| 1001–1100 | 34%/4% | 64% | 49% | 2.77 |
| 2001–2100 | 27%/1% | 56% | 37% | 2.42 |

Table 3: Evaluation of trigram output: unigram subtuples

work, much of the noise found in the lists comes from verbs that primarily appear in the passive—and thus with the preposition 'by'. Simply putting 'by' onto a stoplist would solve the problem, but is unsatisfactory as it rules out true MWVs involving 'by' ("One can *identify* the African swallow *by* its weight-to-wingspan ratio.") More importantly, with actual information about use of the passive, we can get useful information about transitivity.

A related problem is that of dative constructions involving 'to', which also generate a substantial amount of noise in the list. The stoplist solution is even worse here, as there are more legitimate MWVs involving 'to'; in addition, we almost certainly want to distinguish constructions such as "give to", which can undergo dative shift, from "donate to", which cannot.

Another extension we would like to make is to use some of the additional information available in the BLLIP corpus. In particular, the function tags may be of some use in determining whether, for instance, a given prepositional phrase is adjunct or not.

An important continuation of the work would be to extend it to include MWVs with other than just particles. Examples include "take care of", "make mention of", and "file suit". Broadening the scope further, there would seem to be no reason why the techniques presented in this paper couldn't be applied to the collocation problem in general, provided suitable input.

## 6 Conclusion

This work presents two major contributions. First of all, it demonstrates that using parsed input can serve to eliminate a great deal of work in finding the exact target frames—rather than trying to estimate which particles belonged to which verbs, we were able to simply read that information off the tree. Narrowing the search space in this fashion serves to make the searching/ranking algorithm—any algorithm—more efficacious by eliminating spurious entries from the very start. Second, it provides a generalisation and application of the confidence interval algorithm, which proved extremely useful in extracting multi-word verbs of varying lengths, and which should also prove useful in the more general collocation problem.

## References


Alan Agresti. 1990. *Categorical Data Analysis*. John Wiley and Sons, New York.

Alan Agresti. 1996. *An Introduction to Categorical Data Analysis*. John Wiley and Sons, New York.

Ann Bies, Mark Ferguson, Karen Katz, and Robert MacIntyre, 1995. *Bracketing Guidelines for Treebank II Style Penn Treebank Project*, January.

Eugene Charniak, Don Blaheta, Niyu Ge, Keith Hall, John Hale, and Mark Johnson. 1999. Bllip 1987-89 wsj corpus release 1. LDC corpus LDC2000T43.

Kenneth Ward Church and Patrick Hanks. 1989. Word association norms, mutual information and lexicography. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, pages 76–83.

D.A. Cruse. 1986. *Lexical semantics*. Cambridge University Press.

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

Leo A. Goodman. 1970. The multivariate analysis of qualitative data: Interactions among multiple classifications. *Journal of the American Statistical Association*, 65:226–256.

| | Phrasality | Transitivity | Opacity | Goodness of collocation |
|---|---|---|---|---|
| 2-word verbs | 88% | 59% | 74% | 26%; 2.4 avg diff |
| 3-word verbs | 76%/86% | 61% | 67% | 27%; 2.0 avg diff |

Table 4: Interannotator agreement

Myles Hollander and Douglas A. Wolfe. 1999. *Nonparametric statistical methods.* J. Wiley, New York.

Mark Johnson. 2001. Trading recall for precision with confidence sets. Brown University Tech Report.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of statistical natural language processing.* MIT Press.

M.F. Porter. 1980. An algorithm for suffix stripping. *Program*, 14(3).

Randolph Quirk, Sidney Greenbaum, Geoffrey Leach, and Jan Svartvik. 1985. *A comprehensive grammar of the English language.* Longman.