Statistical Measures for Characterising MWE

Ismail El Maarouf and Michael Oakes **WG1**

A large number of statistical measures exist which measure the collocational strength of MWEs, particularly those which are characterised by two main words (Pecina, 2008). Such measures of collocational strength are useful for discovering new pairs of collocates in corpora. One example of a measure of collocational strength is Pointwise Mutual Information (PMI) (Church and Hanks, 1989). In this study we compared 10 MWE which tend to take idiomatic meanings, and found the highest value of PMI was 10.947 for "spill the beans". This is because "spill" and "beans" often occur together in this collocation, but relatively rarely in the corpus as a whole. Another such measure, the simplest, is raw frequency. According to this, the more frequent a collocation, the stronger it is.

In this poster we will also look at statistical measures which have not yet been tested for their ability to discover new collocates, but we have found useful for characterising MWEs containing collocates already found by statistical measures of collocational strength or by prior knowledge of the languages which contain them. Smadja (1993) suggested that collocations should be characterised by whether they are flexible (allowing varying numbers of intervening words between the two words in collocation) or rigid (always having exactly the same number of words between them). To characterise flexibility, we suggest the mean and the standard deviation of the distance in words separating the two collocates, taken over all occurrences of the collocation in the corpus. Thus a rigid collocation would have a standard deviation of 0, while a flexible collocation would have a standard deviation above 0 (the higher the value, the more flexible the collocation).

We also suggest Shannon Diversity (originally developed as a measure of ecological diversity, and equivalent to entropy) as a measure of diversity within a MWE. Does an MWE always consist of exactly the same set of words, or does it take variant forms? The most diverse idiom was "bite your head off", which appeared in exactly this form just 7 out of 40 times. The diversity was mainly due to a) variation in the pronoun, and b) the use of a passive voice variant, as in "heads were bitten off". The least diverse idiom was "spill the beans", which occurred 3 in exactly this form 37 out of 42 times, and thus in variant forms such as "spill some beans" just 5 times.

We also define a measure called Idiomaticity, which is the proportion of times a MWE is found in the corpus where it takes its idiomatic, rather than a literal meaning. The maximum value is 1, such as for the phrase "bite the bullet" which was found only with its idiomatic meaning in the British National Corpus. The phrase with the least idiomaticity of those we examined was "kick the bucket", which was idiomatic in only 5 out of 21 instances, giving an idiomaticity of 5/21 = 0.238. "bite back" is an interesting case, because it can take either of two idiomatic meanings: to restrain oneself, as in "biting back remarks", or to take revenge as in "bite back at bad behaviour". It took a literal meaning just once, in "the rats … bite back". He we could define the idiomaticity of each meaning separately. Since there were 64 instances of the first meaning out of a total of 88 instances of "bite back", the idiomaticity of the first meaning would be 64 / 88 = 0.681; for the second meaning it would be 23 / 88 = 0.261. The data in the table for "bite back" is for both idiomatic meanings combined.

The longest MWEs on average were variants of the idiom "bite the hand that feeds you". A corpus search for sentences containing both "bite" and "hand" yielded a large number of examples, where

the words “bite” and “hand” were generally not related. To identify idioms of the form we wanted, we needed to specify a third component, namely that the hand that is bitten bestows some form of benefit. In 7 cases this was “feeds”, but in other cases the benefit was more elaborately described, as in “the hand with which they had so kindly offered freedom had been bitten”. The shortest MWE on average was “bite back”, which in every case but four consisted of just these two words. The most flexible MWE, shown by the high standard deviation of its length, was “kick the bucket”. This was almost entirely due to a single instance “Arthur kicked the detonator of the bomb, and consequently the bucket”. The most rigid MWE, shown by its low standard deviation of the length, was “bite the bullet”. This was because in almost every case (29/36) the exact phrase “bit* the bullet” was used, just twice it was one word shorter in “bite bullets”, and in just 5 cases it was one word longer (“biting on the bullet” or “biting the ideological bullet”).

Examples of MWE in the table which we have not mentioned (because their values by each of our criteria were unexceptional) were “which way the wind blows”, “quake in his boots”, and “clutch at straws”. The idioms in the table were characterised by the presence of both a main noun and verb, except for “which way the wind blows” which also required the presence of the word “way”, “bite the hand that feeds” also needed some sort of benefit, and “bite your head off” also needed the preposition “off”. Grammatical variants (such as tense or number) of the main noun and main verb were regarded as equivalent, but variants of the intervening words were regarded as distinct.

idiom	Freq	PMI	idiomaticity	entropy	Mean length	Standard deviation
[way][wind][blows]	20	10.663	0.645	2.158	3.8	0.523
[shoe/boot] [quake/shiver/shake]	13	5.043	1	3.393	3.333	1.155
[straw][grasp/clutch]	33	9.865	0.892	2.849	2.061	1.560
[bean][spill]	40	10.947	0.952	0.503	1.775	1.000
[bucket][kick]	5	8.647	0.238	0.721	3.4	3.130
[bullet][bite]	36	10.484	1	1.069	1.944	0.791
[hand][bite][BENEFIT]	15	5.584	1	2.463	6.071	2.073
[bug][bite]	16	10.589	0.696	3.406	2.125	2.802
[back][bite]	87	5.914	0.977	1.150	1.057	0.279
[head][bite][off]	33	6.009	0.775	4.074	1.030	1.960

To determine whether these measures were independent of each other or whether one might act as a predictor for another, Pearson’s correlation coefficient was found for each pair of measures using the 10 idioms in the table. The only statistically significant correlations were between frequency and mean length ($p = 0.0341$, $cor = -0.670$) and between frequency and the standard deviation of the length ($p = 0.315$, $cor = -0.677$). Thus there was a tendency for more frequent idioms to be shorter and more rigid in their structures. One other correlation was almost significant, that between idiomaticity and standard deviation of length ($p = 0.055$, $cor = -0.622$). Here there is a tendency for MWEs which more frequently take an idiomatic meaning to be more rigid in their structures. There was no correlation between any of the measures in the table with either of the measures of collocational strength, frequency and PMI. This suggests that the new measures of idiomaticity, entropy, mean length and standard deviation of length may not be useful for discovering new MWE, but as we have shown, are useful for describing the characteristics of MWE once discovered.