Automatic extraction of MWEs for the Pattern Dictionary of English Verbs.

Patrick Hanks, Ismail El Maarouf, Michael Oakes.

RIILP, University of Wolverhampton, England.

In this chapter we describe the work of the Corpus Pattern Analysis project, led by Prof. Patrick Hanks, in the development of the Pattern Dictionary of English Verbs (PDEV), which lists all the senses of each verb as defined by the linguistic patterns in which they occur. For example, in the Pattern Dictionary of English Verbs (PDEV), (http://www.pdev.org.uk) the sense of "blow" in "blow your nose" is stored in the pattern *[Human] blow {nose}* while the sense of "blow" in "the wind blows" is represented by the pattern *[Wind | Vapour | Dust] blow [No object] [Adverb of direction]*.

We propose the use of statistics developed initially by Church and Hanks (1989) to speed up the discovery of these verb patterns, and once found, for annotating them in PDEV. The work is relevant to PARSEME, since statistical measures of collocational strength have often been used to detect MWEs – sets of words which collocate strongly are more likely to be meaningful MWEs. Our approach to finding such MWEs is hybrid, since we first find candidate sets of syntactically-related words using the Stanford Parser (Klein and Manning, 2003), and then we use statistics to rank these candidate MWEs in order of their collocational strength. Our corpus was a 50 million-word subset of the British National Corpus.

Collocational Strength

In psycholinguistics, "word association" means for example that subjects think of a term such as "nurse" more quickly after the stimulus of a related term such as "doctor". Church and Hanks (1989) redefined "word association" in terms of objective statistical measures designed to show whether a pair of words are found together in text more frequently than one would expect by chance. PMI between word x and word y is given by the formula $I(x,y) = \log_2 P(x,y) / P(x).P(y)$, where P(x,y) is the probability of the two words occurring in a common context (such as a span of 5 words, or in subject-object relation), while P(x) and P(y) are the probabilities of finding words x and y respectively anywhere in the corpus. PMI which is positive if the two words tend to co-occur, 0 if they occur together as often as one would expect by chance, and less than 0 if they are in complementary distribution (Church and Hanks, 1989). PMI was used by Church and Hanks to examine the content word collocates of the verb "shower", which were found to include "abuse", "accolades", "affection", "applause", "arrows" and "attention". Human examination of these lists is needed to identify the "seed" members of categories with which the verb can occur, such as speech acts and physical objects, giving at least two senses of the verb (Hanks, 2012). While PMI is useful for finding the strength of association between just two words, it can be extended to produce association measures for three words (Van de Cruys, 2012). Two variants suggested by Van de Cruys are Specific Correlation (SC) and Specific Interaction Information (SII), as shown in the following formulas:

$$SC(x,y,z) = \log_2 \frac{p(x,y,z)}{p(x)\,p(y)\,p(z)}$$

$$SII(x,y,z) = \log_2 \frac{p(x,y)\,p(y,z)\,p(x,z)}{p(x)\,p(y)\,p(z)\,p(x,y,z)}$$

Highly scoring SVO triples according to the SC measure were "Value added tax", "glazed UPVC window", "maximum branching ratio" and "stamped addressed envelope". The highest scoring triples for both SC and SII were compared against PDEV's manually-prepared idiom list, and it was found that SC captured slightly more of these idioms.

Flexibility, diversity and idiomaticity of collocations

Smadja (1993) recommends that collocations should not only be measured by their strength, such as by using the z-score, but also by their flexibility. This can be done by finding the mean of the relative distances between two words, and the *spread* of each collocation, which is the standard deviation of the relative distances between the two words. High spread would indicate a flexible or semantic, rather than a rigid, lexical collocation. In a study of David Wyllie's English translation of Kafka's *Metamorphosis*, Oakes (2012) found that *stuck fast* and *office assistant* had mean inter-word distances of 1 with a standard deviation of 0. This showed that in this particular text, they were completely fixed collocations where the first word was always immediately followed by the second. Conversely, *collection* and *samples* had a mean distance of 2.5 with a standard deviation of 0.25. This collocation was a little more flexible, occurring both as *collection of samples* and *collection of textile samples*. *Mr. Samsa* had a mean distance of 1.17 and a standard deviation of 0.32. This is because it usually appeared as *Mr. Samsa* with no intervening words, but sometimes as *Mr. and Mrs. Samsa*.

Another way of looking at the flexibility of a collocation is by measuring the diversity of surface forms found for that collocation. A rigid collocation, where all found examples are identical in form and length, has very low diversity, while a collocation which has many surface forms has much higher diversity. One measure of diversity, popular in ecological studies, is Shannon's diversity index, which is equivalent to entropy in information theory, and given by the formula:

$$E = -\sum_{i=1}^{N} p_{i \log_2} p_i$$

E is entropy, N is the number of different surface forms found for the collocation, i refers to each surface form in turn, and p_i is the proportion of all surface forms made up of the surface form currently under consideration. The choice of logarithms to the base 2 ensures that the units of diversity are bits. The minimum value of diversity (when all the examples of a phrase or idiom are identical) is 0, while the maximum value (when all the examples occur in different forms) is the logarithm to the base 2 of the number of examples found. For standard deviation, the minimum value when all the examples are identical length is 0, and there is no theoretical upper limit. In our experiments on the 2 million-word subset of the BNC corpus, we found that the phrase "bite the bullet" was maximally rigid, as it occurred all 9 times in exactly that form. Thus the standard deviation of the collocation length was 0, and its diversity was also 0. In contrast, the phrase "bitten by the … bug" was extremely flexible, occurring all 6 times in different forms such as "bitten by the travel bug", "bitten by the London bug", and "bitten by the bug of the ocean floor". The standard deviation of lengths (0.48) was relatively small, reflecting that in all cases but one the variation consisted of the insertion of a single word, but the diversity index was its maximum value for a set of 6 examples, $\log_2(6) = 2.585$.

The results for "bite" were borne out when the experiment was repeated on a larger corpus, the entire BNC. There were 25 occurrences of "[bite] by X the bug" altogether, where "[bite]" stands for any grammatical variant of "bite", such as "bitten", and X stands for any number (possibly zero) of intervening words. 14 of these were idiomatic; 4 of the non-idiomatic examples were variants of the farewell "sleep tight, don't let the bed bugs bite", and 5 were literal as in "I've been bitten by bugs in a hooker's bed"; 2 were difficult to determine from the context. Another measure for characterising MWEs is idiomaticity, which is the ratio of the number of idiomatic occurrences of the phrase divided by the total number of occurrences of the word = 16 / 25 = 0.64. Of the idiomatic examples, almost all were unique, such as "bitten by the travel bug" – the other "bugs" included puppy love, acting,

the ocean floor, racing, flower pressing, showbiz, London, newspaper, gold, drama, golf and flying. On 3 of these occasions the nature of the bug did not appear between “bitten” and “bug”, which were simply connected as “bitten by the bug”. The Shannon diversity, resulting from 3 similar and 13 unique occurrences, had a very high value of 3.708. In terms of flexibility, the mean distance between “[bite]” and “bug]” was 3.0, with a high standard deviation of 2.07. This was because a number of cases, such as “the acting bug really bit me” used the passive voice, so “bug” appeared before “bit”. Also influencing flexibility was the fact that even in the active voice, the number of intervening words could vary.

“[bite] the bullet” occurred in 33 sentences altogether, there were no literal examples at all, but “bite the bullet” appeared as the name of both a racehorse and a pop song. Of the other 31 examples, the vast majority (27) were exactly in the form “[bite] the bullet”, the remainder being in the forms “bit the ideological bullet” (2); reversed as in a “harder bullet to bite” (1) and a statement by President Bush about an opponent: “I bite bullets, he bites nails”. Idiomaticity was thus at a theoretical maximum value of 31/31 = 1.00. The collocation was rather rigid, with a mean separation between “[bite]” and bullet of 1.97 (very close to 2), and a fairly low standard deviation of sd = 0.80. Diversity was also fairly low at 0.748.

We then examined these measures for the idioms “kick the bucket” and “spill the beans”. In the BNC, “[kick] the bucket” has 24 occurrences, although 4 were discounted as “kick” and “bucket” appeared in separate clauses. Another 5 were from a linguistic discussion of the phrase, as in “notice ‘kick the bucket’ appears as a verb phrase”. Only 6 were idiomatic, in the sense of “to die”: 5 of these were in the exact form “kicked the bucket”, while the other had a sequence of 9 words between “kicked” and “bucket”, in “Arthur kicked the detonator of the bomb, and consequently the bucket”. This gave a mean separation of 3.75 and a high standard deviation of 3.5, and a modest diversity of 0.811. However, these results were biased by a small sample size and a single creative use of language. This left 8 literal examples of the phrase, as in “leaving his bucket to be kicked over by the cow”. Thus idiomaticity was 6 /(9 + 6) = 0.40. In contrast, the phrase “spill the beans”, found 43 times overall in the BNC, was almost always (40 times) found in the idiomatic sense of “reveal a secret”. The only exceptions were when the phrase was used as the title of a book, “A style guide to the New Age called ‘spilling the beans’”, and a television programme “Superchefs spill the beans”, where the phrase “spill the beans” takes both the literal and the figurative sense at the same time. The phrase was used just once in its purely literal sense, where a guest house owner was dreading “a dozen or more children spilling their beans, wetting the beds, hoarding old crusts”. Thus idiomaticity was very high = 40 / 41 = 0.98. Of the 40 idiomatic cases, the vast majority were in the exact form “[spill] the beans” (37); 2 were in the passive voice (“when the beans are spilled”) and (“the beans have been spilled”), and just one replaced “the” with “a few”: “he spilt a few beans”. The means separation was 1.8, the flexibility as measured by the standard deviation was 1.02, and diversity as measured by entropy was a lowish value of 0.503. According to these results, “spill the beans” is more idiomatic, less flexible and slightly less diverse than “kick the bucket”. These findings are in stark contrast with the fact that MWEs like “spill the beans” are often reportedly more flexible than the relatively well behaved “kick the bucket”.

The statistical criteria described in this chapter can be used not as an alternative, but additionally to symbolic MWE classification criteria. The symbolic criteria will define the constraints to which the MWE must conform, then the MWE as defined can be described numerically with measures of collocational strength, flexibility, diversity and idiomaticity.

Kenneth W. Church and Patrick Hanks. Word Association Norms, Mutual Information and Lexicography. 1989. *27th ACL:* 78-83.

Patrick Hanks, *How People Use Words to Make Meanings. Semantic Types meet Valencies.* 2012. In J. Thomas and A. Boulton (eds.), Input Process and Product: Developments in Teaching and Language Corpora. Masaryk University Press.

Dan Klein and Christopher D. Manning. 2003. Accurate Unlexicalized Parsing. *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, pp. 423-430.

Michael P. Oakes.. Describing a Translational Corpus. In: Oakes, M. P. and Ji, M., Quantitative Methods in Corpus-Based Translation Studies, Amsterdam: John Benjamins, 2012: 115-148.

Frank Smadja. Retrieving collocations from text: Xtract. 1993. Computational Linguistics 19: 143-177.

Wikipedia. Diversity Index. http://en.wikipedia.org/wiki/Diversity_index

Tim Van de Cruys. Two Multivariate Generalizations of Pointwise Mutual Information. disco 2011, 24 June 2011, Portland, OR