Statistical Measures to characterise MWEs involving "mordre" in French or "bite" in English.

Ismail El Maarouf and Michael Oakes.

Oxford University Press and University of Wolverhampton.

WG3

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. In this paper we will 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. Church and Hanks (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, considering all forms of the lemmas of the collocates. 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). For example, the corresponding phrases "mordre la poussière" and "bite the dust" both have standard deviations for their lengths of 0, since in the BNC and Frtenten corpora the verb is always exactly 2 words before the noun. We also suggest Shannon Diversity (originally developed as a measure of ecological diversity) 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? A phrase like "bite the bullet" in English always (in the BNC corpus) occurs as exactly these three words, so its diversity is 0, while "bitten by the bug" takes many forms: "bitten by the acting bug", "bitten by the travel bug", "bitten by the golf bug", and so on. The diversity of "bitten by the bug" is close to its maximum theoretical value of the logarithm to the base 2 of the number of examples in the corpus. Its French counterpart, "mordue de" is also highly diverse, as in the examples "mordue des nuitées en famille sous la tente" (fanatical about nights camping with the family), "mordus des jeux on ligne" (addicted to on-line games) and "mordue d'esperanto" (bitten by the Esperanto bug). The pattern "[[Human]] se mord {les doigts}" rarely takes its literal meaning in French, more often standing for "a person experiencing a bitter time for his past actions". It usually occurs in the Frtenten corpus as "mordre les doigts", but sometimes as "mord encore les doigts" (bites his fingers again), "mordrait un peu souvent les doigts" (bit his fingers a bit often) and other variants. This gave a mean and standard deviation of the lengths of 1.19 and 0.15, and a Shannon Diversity of 1.08.

We applied the statistical measures to all the French and English idioms in the FrTenTen and BNC50 corpora containing "mordre" and "to bite" respectively. The results revealed that idioms come in a variety of forms, and have diverse properties on the scales of spread, flexibility, and diversity. In our experiments, we found that the idiom "bite the bullet" was maximally rigid, as it occurred all 9 times in exactly that form, with standard deviation and entropy both equal to 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 (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 6 examples, log2(6) = 2.58. In French, one pattern of “mordre” (“le poisson mord à hameçon/l’appât’) comes in three different forms which have equivalent meaning (“to take the bait”), and have therefore been studied separately: the collocates “poisson”, “hameçon”, and “appât” were never found together. For each of them, independently of the spread (mean length) or its frequency, the idiom was always fixed (standard deviation = 0) and never took variants (Entropy = 0). The idiom “[[Human]] se mord les doigts” usually occurred as “mordre les doigts”, but sometimes as “mord encore les doigts” (“bites his fingers again”), “mordrait un peu trop souvent les doigts” (“bit his fingers a bit too often”) and other variants. This gave a mean, standard deviation, and entropy of 2.3, 0.9, and 0.84 respectively. The corresponding phrases “mordre la poussière” and “bite the dust” both have standard deviations and entropy close to 0, since, in both corpora, they allow very little variation.

Overall, the measures of standard deviation and entropy seem to coincide with intuitions about the rigidity of idioms, the scores being most of the time either equal to zero (maximally rigid and not diverse), or below 1, which still qualifies as low. An important issue not touched upon in this paper, is the idiomaticity of an expression, that is the proportion of uses that do not have a literal interpretation. In a pilot study, we found that not all instances of a MWE defined as idiomatic, take an idiomatic reading in context. For example, we found that only 6 of the 24 expressions formed with the boundary words “kick” and “bucket”, and in appropriate syntactic relation, had an idiomatic reading.

MWE are challenging not for only second language speakers, but also for MT systems. For example, "bite one's fingers" and its apparent French translation "se mordre les doigts" are in stark idiomaticity contrast. While "bite one's fingers" was always found to be literal (5 cases), all instances of “se mordre les doigts” (21) were found to be idiomatic. Systems unaware of this will tend to make two mistakes (as can be checked with Google Translate): when translating from French to English, they will fail to translate the figurative meaning of "se mordre les doigts" with an equivalent idiom like "kick oneself". From English to French, they will fail to translate the literal meaning of "bite one's fingers" and translate it with the frequent idiomatic sequence "se mordre les doigts". For the verbs “mordre” and “bite”, we have shown that the measures of mean and standard deviation of length, Shannon Diversity, and idiomaticity (proportion of occurrences which are idiomatic) give intuitively reasonable results. We propose these measures as parameters in an MT system.

Pavel Pecina. 2008. Lexical Association Measures: Collocation Extraction. Ph.D. Thesis, Charles University in Prague.

Frank Smadja. 1993. Retrieving Collocations from Text: Xtract. Computational Linguistics 19: 143-177.