Collocationality (and how to measure it)

Adam Kilgarriff
Lexical Computing Ltd
Brighton, UK
adam@lexmasterclass.com

Abstract
Collocation is increasingly recognised as a central aspect of language, a fact that English learners' dictionaries have responded to extensively. Statistical measures for identifying collocations in large corpora are now well-established. We move on to a further issue: which words have a particularly strong tendency to occur in collocations, or are most 'collocational', and thereby merit having their collocates shown in dictionaries. We propose a measure of collocationality based on entropy, as defined in Information Theory. We describe experiments to find the most collocational words in the British National Corpus, present results with the most collocational nouns and verbs in relation to the grammatical relation OBJECT, and compare the results to collocational words identified in Macmillan English Dictionary for Advanced Learners.

1 Introduction

As Firth pointed out, "you shall know a word by the company it keeps" (Firth 1968:179). The dictum has been taken to heart, in linguistics (Hoey 2005), psycholinguistics (Wray forthcoming), language teaching (Carter and McCarthy 1988) and lexicography (Hanks 2002).

As we gain access to large corpora, so it becomes possible to use statistical methods to automatically identify collocations. This line of "lexical statistics" research was inaugurated by Church and Hanks's 1989 paper which introduced Mutual Information (MI), a measure from Information Theory (Shannon and Weaver 1963), as a statistic for measuring how closely related two words were. Since then there have been both a number of papers refining and comparing different statistics (e.g. Dunning 1993, Krenn and Evert 2003), and very widespread use of MI and related statistics, particularly as they are computed, and collocation lists of highest-scoring items presented, in all the widely used corpus query tools.

MI and related measures are measures of association. They assess how noteworthy the association is between two words. In this paper we address a slightly different question: which words have a strong tendency to occur in collocations? Which words are very "collocational"?

While the association measures are more basic, the issue of collocationality arises in a number of contexts. One is language teaching: what are good words to use, to start teaching about collocationality? Another is lexicographic. Dictionaries such as the Macmillan English Dictionary for Advanced Learners (2002) and the Cambridge Advanced Learner's Dictionary

(2005) present brief accounts of collocations for words where collocation is a particularly salient theme. Which words should these accounts appear for?

2 Entropy

A word that is very 'collocational' is one which has a strong tendency to appear with particular words, rather than appearing freely with large numbers of words. This theme is captured mathematically by 'entropy', again from Information Theory. Entropy is defined over a probability distribution, and states how much information there is in that distribution. (A *probability distribution* is a set of possible outcomes, and the probability of each, so for an unbiased coin, the probability distribution is

Heads, 0.5
Tails, 0.5)

The entropy is defined as the sum, across all possible outcomes, of the product of the probability and the log (to base 2) of the probability (which is a negative quantity, so we then change the sign to positive):

- $p(x).log(p(x))$

2.1 Probabilities, proportions

How can a word's collocations be modeled as a probability distribution? We take a straightforward approach. We view all the words that occur with a node word as the set of possibilities. We then count how often each of them occurs with the nodeword, in a corpus. We can then estimate the probability for each collocate, as the proportion of the complete set of occurrences of the node word with some collocate or other, that occurred with this particular collocate. Technically, this is a *maximum likelihood estimate* (MLE) for the probability of the collocate, given the data.[1]

2.2 Grammatical relations

Next we clarify what we mean by the collocate occurring "with" the nodeword.

In the linguistics literature, relations between base and collocates are generally grammatical. Prototypical collocations associate a base noun with the verb it is object of (*pay attention*) or a base noun with an adjective that modifies it (*bright idea*). In dictionaries such as the Oxford Collocations Dictionary (OCD; 2002), collocates are divided according to the grammatical relation they stand in to the base: in noun entries, OCD typically first lists adjectives that modify the noun, then verbs that the noun is object of, then verbs that the noun is subject of, then prepositions, then phrases.

Our experience with corpus-derived data also demonstrates the usefulness of grammatical relations as an organizing principle. We use the Sketch Engine (Kilgarriff et al 2004) and

[1] While MLEs are not good estimates for probabilities when based on low counts, and low-frequency collocates will be very common, we ask readers to look to the results to assess whether these considerations cause problems here.

note the merit as argued in Kilgarriff and Rundell (2002) of having a separate list of high-salience collocates for each grammatical relation. This both classifies collocates and provides a method for filtering out the high levels of noise that are otherwise typical of collocate lists.

In this report, we treat each grammatical relation separately. Thus we aim to identify the most collocational items with respect to a particular grammatical relation, for example, the most collocational nouns with respect to the verbs they are object of. We can envisage a number of ways in which different lists might be merged, but leave that for further work.

So: a collocate occurs "with" a nodeword if it occurs in the specified grammatical relation to the nodeword.

3 Experiment

We used the British National Corpus[2] – a 100 million word corpus of spoken and written British English, covering a wide range of text types – as loaded into the Sketch Engine. The version of the BNC we used was lemmatized, so we could treat grammatical relations as holding between lemmas ("take") rather than word forms ("take" or "took" or "takes" or "taken" or "taking") and part-of-speech-tagged. The Sketch Engine supports shallow parsing, and we used this parsing facility to identify all instances of triples of the form <grammatical-relation, nodeword, word2> in the corpus.

For illustration we use the relation 'object', where the nodeword is a noun and the second word is a verb. We identified, for each noun, what verbs it occurred as object of, and how often. For the noun *advantage* we identified the verbs below, which we treat as the population of possible outcomes. Our estimates of probabilities are then, for each verb, the number of times that *advantage* has that verb as object, divided by the total number of times that a verb was identified which had *advantage* as its object. (This means that the sum of the proportions, in the third column below, is 1, as required for a probability distribution.)

Verb	Freq	probability (freq/3730)	Log	-(prob x log)
Take	2084	.5587	-0.84	.469
Gain	131	.0351	-4.83	.169
Offer	117	.0314	-4.99	.157
See	110	.0295	-5.08	.150
Enjoy	67	.0180	-5.79	.104
Obtain	58	.0155	-6.01	.093
...	...	...	...	...
Clarify	1	.000268	-11.86	0.0031
...	...	...	...	...
Total	**3730**	**1.000**		**3.909**

Table 1. Calculation of entropy for *advantage (object relation)*.

[2] http://www.natcorp.ox.ac.uk

We perform the calculation as above: the entropy of the noun *advantage,* in relation to the object relation, given BNC data, is thus 3.909.

We also calculate entropy for all other nouns in relation to 'object'. (We limit the data set to nouns found as objects more than fifty times in the BNC, using the Sketch Engine's lemmatiser, parser, etc.)

Plotting entropy against the frequency of the noun (occurring with an identifiable object, eg 3730 for *advantage)* gives us the graph in Figure 1. Each cross represents an individual noun. Note that the frequency axis is logarithmic.

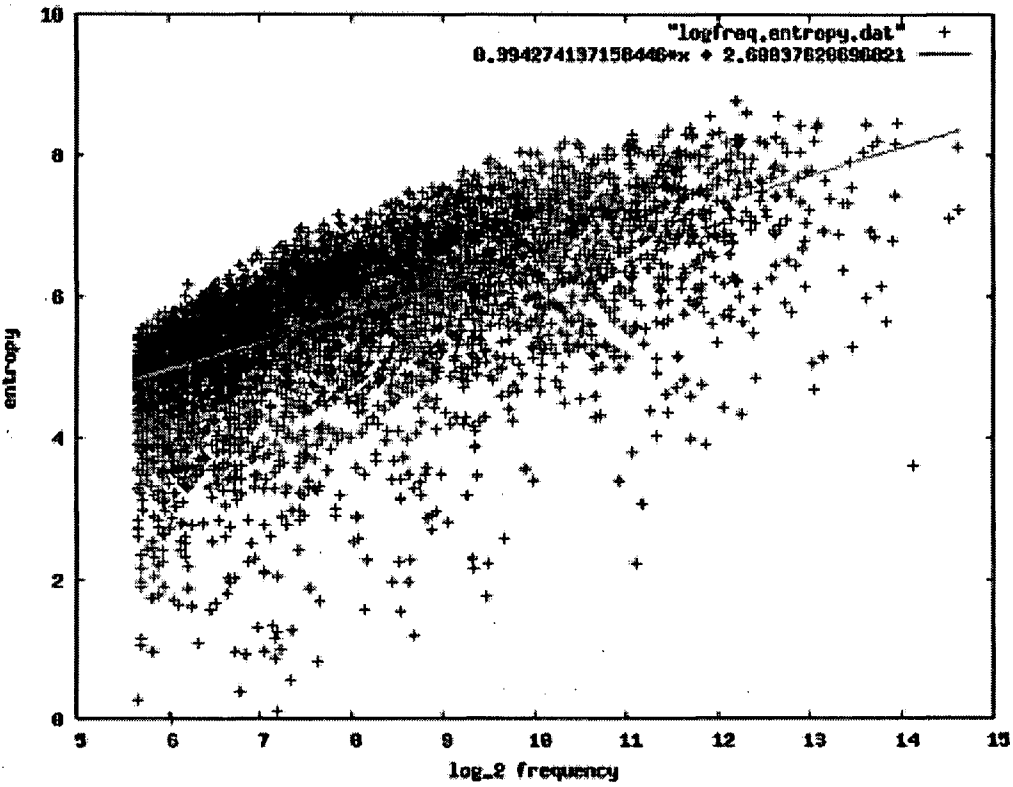

Figure 1. Graph of entropy vs frequency for *object-of* relation for 5738 nouns in the BNC

It is evident from the graph that entropy tends to increase with frequency. As can be seen, the lowest-entropy items are low-frequency items. It is a characteristic of entropy that it tends to increase with the number of possible outcomes (here, 'possible outcomes' are "verbs that the noun has occurred as object of", so more frequent nouns tend to have more possible outcomes). We consider collocationality a "frequency-neutral" term: that is, we would like to say that common words do not intrinsically have any stronger likelihood to be highly collocational than rare words. Thus, in order to use entropy as the basis for a measure

of collocationality, we should first normalize it, to take away frequency effects. There are various ways this could be done. We use the simplest. We take the minimum-regression straight line through the graph, that is, the line which points on the graph are, on average, closest to. The normalized value is then the distance from this line (which is shown on the graph).

4 Results

We present below the 100 "most collocational" items for nouns (with respect to the verbs they are objects of) and verbs (with respect to their object-nouns) ordered by frequency. Frequency ordering is useful as the higher-frequency words often present different patterns to the lower-frequency ones. (Numbers in brackets give the number of data points each result is based on.) We discuss nouns/objects in some detail: for verbs/objects, we make only brief observations.

4.1 Nouns/objects

place (17881), attention (8476), door (8426), care (4884), step (4277), advantage (3730), rise (3334), attempt (2825), impression (2596), notice (2462), chapter (2318), mistake (2205), breath (2140), hold (1949), birth (1016), living (953), indication (812), tribute (720), debut (714), button (661), eyebrow (649), anniversary (637), mention (615), glimpse (531), suicide (486), toll (472), refuge (470), spokesman (453), sigh (436), birthday (429), wicket (412), appendix (410), pardon (399), precaution (396), temptation (374), goodbye (372), fuss (366), resemblance (350), goodness (288), precedence (285), havoc (270), tennis (266), comeback (260), farewell (228), prominence (228), go-ahead (202), sip (198), accountancy (188), climax (173), nod (172), brunt (163), headway (161), fancy (157), damn (151), plunge (147), credence (146), amends (146), piss (145), inroad (145), sway (142), communiqué (140), crying (133), para (133), overdose (132), heed (127), toss (124), centenary (121), detour (117), sae (115), hang (110), shrug (107), stir (106), save (103), gamble (101), cholangitis (100), chess (92), stroll (89), twinge (80), papillum (76), virginity (75), errand (75), gauntlet (73), gulp (69), bluff (67), swig (66), robemaker (62), cool (61), doorbell (60), glossary (59), shrift (58), keep (57), spokeswoman (57), grab (56), snort (52), quarter-final (52), fray (52), esc (52), cropper (52), mickey (51)

Many of these nouns occur with high frequency in support-verb constructions: thus events *take place*, we *pay attention* and *take care*, *steps* and *advantage* (of situations). *Door* does not have one support verb but, rather, seven strong collocates: as well as *opening* doors, we *close*, *shut*, *lock*, *slam*, *push* and *unlock* them. (Checking is easily undertaken, as the data is the same as the data used to prepare the word sketch for the noun, so the word sketch, which is hyperlinked to the relevant KWIC concordance, can be examined, e.g. at http://www.sketchengine.co.uk.)

Looking at the lower-frequent items towards the end of the list, we find idioms where the noun rarely or never occurs outside the idiom: *take the mickey*, *come a cropper*, *enter* (or *join*) *the fray*, *make headway*, *bear the brunt* (of), *take the piss*, *beg pardon* and *give the go-*

ahead. Shrift (which is nearly always *short)* is *given, got* and *received.* We see a formulaic aspect of sports journalism in teams usually *reaching quarter-finals.*

A little noise must be set aside. *Appendix, chapter, page, glossary* and *para* result from the verb *see* as in "see chapter 5". Sae features because people are often asked to "send SAE" (self-addressed envelope), and SAE rarely occurs anywhere else.[3] *Accountancy* results from self-references in one of the BNC's sources, the magazine "Accountancy", which often says, eg, "See ACCOUNTANCY, Jan 9 1992, p 21". *Esc* results from pages of computer manuals which often say "Tap ESC for ESCAPE". (Corpus statistics of this kind aim to find surprising facts about the language, and surprisingness is identified with being much more common than the norm. So the statistics also succeed in finding items that are strikingly common for arbitrary non-linguistic reasons, often to do with the composition, collection and encoding of the corpus. The issue arises throughout corpus linguistics.)

The items which occur largely with one verb, in fairly fixed expressions, will already be specified in good dictionaries. The words that are more lexicographically useful are those where the entropy is low, but the list of strong collocates is long enough and diverse enough for the entry to be a useful locus for some description of collocational behaviour. Around half of the high-frequency, high-collocationality items meet this criterion: *attention (draw, pay, attract, give, focus, turn), care (take, provide, need), impression (give, make, get, create), notice (serve, take, give), breath (catch, draw, take, hold), hold (grab, get, take, catch, keep).*[4]

4.1.1 Macmillan comparison

We identified the 322 nouns for which there were collocation panels in the Macmillan Dictionary, and considered their collocationality scores with respect to the object relation. They had markedly higher scores than a random sample of nouns: eight were in the top hundred, a quarter were amongst the highest-scoring 11%, and two thirds were in the top half. Where a word had a low collocation score, but nonetheless had a Macmillan panel, there are three possible interpretations: the word may have strong collocates in other grammatical relations; the collocation measure may be flawed; or Macmillan's selection may be open to improvement. The motivation for this paper arose following a conversation with Macmillan's editor in which we wondered how to make the selection of words for collocation panels more principled, so we do not believe there are principles to the Macmillan selection which we are overlooking.

4.1.2 Verbs/objects

take (106749), pay (18925), play (17832), raise (15477), spend (15267), open (11362), close (6106), shake (5483), sign (5100), answer (4177), exercise (3265), speak (3013), solve (2555),

[3] Capitalisation would ideally be taken into account, although we note that capitalisation in most corpora – even carefully edited ones such as the BNC – is an unreliable clue to linguistics status. Words may be fully capitalised because they are at the beginnings of stories, in headings, or for emphasis, and there are also interactions with the POS-tagger which uses capitalisation as part of its evaluation of whether a word is a proper name. Proper names are excluded from this list.

[4] Here we list a verb if it accounts for over 5% of the data. Diversity was harder to evaluate, and *impression* is a marginal case as the verbs are not so diverse.

score (2495), live (2201), waste (2091), thank (1926), pose (1897), fulfil (1885), wait (1768), shut (1675), last (1521), incur (1365), research (1072), devote (1025), age (1009), exert (966), bite (919), park (836), beg (739), slam (634), sip (574), narrow (540), levy (450), nod (433), part (425), adjourn (424), pave (420), clasp (411), ratify (391), reap (376), bridge (337), shrug (324), enlist (322), clench (313), bow (303), wage (299), clap (256), redress (248), dial (232), retrace (205), poll (202), cock (200), coin (194), comb (193), purse (191), grit (170), stake (169), allay (167), wring (157), wag (154), peacekeep (151), fell (147), incline (139), wreak (138), ruffle (136), wrinkle (134), preheat (134), adduce (133), broach (122), foot (121), hunch (109), blink (103), bide (103), disobey (99), whet (89), sclerose (86), jog (85), buck (85), moisten (81), jumble (81), recharge (81), wuther (81), overstep (74), scroll (74), crane (74), hazard (70), mince (66), pervert (65), elapse (60), hesitate (60), grope (59), elbow (57), re-run (57), transact (55), contort (55), redouble (55), immunise (53), pry (52)

It is interesting to see *take* in the list, and this clearly reflects its common role as a support verb. It has a long list of very strong collocates, including many of the items in the nouns list above. Some other items in this group are corollaries of nouns in the noun/objects list (*slam doors, beg pardon).* Some relate to past and present participles taking adjectival roles, rather than finite verbs (*peacekeeping force/troops/operation, redoubled efforts, sclerosing cholangitis)* and some items, like this last, relate to specialist documents in the corpus. *Shake (hand, head, fist), wag (tail, finger, dog)*, clap (hands, eyes) and clasp (hands, eyes, fngers) form an intriguing group.

5 Discussion

The measure captures some aspects of lexicographically interesting collocationality. It has the additional merit of being based on well-understood mathematics, from Information Theory.

It gives greatest weight to those bases which occur predominantly with just one strong collocate: while this provides one useful list, another will focus on words with a small number of strong collocates.

The measure makes no acknowledgement of the frequency of the collocate. Collocates which, in the corpus at large, are lower-frequency make more striking collocates, as is explored extensively in relation to measures of collocation strength: it is as yet unclear whether this should play a role in a measure for collocationality. Intuitively, it seems appropriate that *enlist help* contributes more to the collocationality score for *help* than *provide help*, even though it is less frequent. Entropy does not capture the intuition.

Collocationality as explored in this paper is grammatical-relation-specific. It may be useful to bring together data from different grammatical relations to provide a unified collocationality profile for a word.

Each of these proposed extensions will lead away from direct use of entropy as the underlying mathematical idea. The approach taken here is to say that, while this may be inevitable, it is good to start from as simple mathematics as possible.

We have presented a first corpus-driven, implemented account of collocationality, and we hope that it will stimulate further work on this aspect of the lexis of the language.

References

A. Dictionaries

Cambridge Advanced Learner's Dictionary, 2nd edition 2005. CUP.
Macmillan English Dictionary for Advanced Learners. 2002. Macmillan.
Oxford Collocations Dictionary. 2002. OUP.

B. Other Literature

Carter, R., McCarthy, M. (1988), *Vocabulary and Language Teaching*, Longman.

Church, K., Hanks, P. (1989), 'Word Association Norms, Mutual Information, and Lexicography', in *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*; reprinted in *Computational Linguistics*, Spring 1990.

Dunning, T. (1993), 'Accurate models for the statistics of surprise and coincidence', *Computational Linguistics*, 19(1), pp. 61-74.

Evert, S., Krenn, B. (2001), 'Methods for the qualitative evaluation of lexical association measures', in *Proc. 39th Meeting of the Association for Computational Linguistics*, Toulouse, France, pp. 188-193.

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.

Hanks, P. (2002), 'Mapping Meaning onto Use', in Corréard, M-H. (ed.) *Lexicography and Natural Language Processing: a Festschrift in honour of B. T. S. Atkins.* Euralex.

Hoey, M. (2005), *Lexical Priming: A new theory of words and language.* Routledge.

Kilgarriff, A., Rundell, M. (2002), 'Lexical profiling software and its lexicographic applications – a case study', in *Proc EURALEX*, Copenhagen, August, pp. 807-818.

Kilgarriff, A., Rychly, P., Smrz, P. and Tugwell, D. (2004), 'The Sketch Engine', Proc. Euralex, Lorient, France, July, pp. 105-116.

Shannon, C., Weaver, W. (1963), *The Mathematical Theory of Communication.* Univ of Illinois Press.

Wray, A. (2006), 'Formulaic Language', in Brown, K. (ed.) *Encyclopedia of Language and Linguistics.* 2nd edition, Oxford, Elsevier.