Chapter 4

Flexibility of multiword expressions and Corpus Pattern Analysis

Patrick Hanks
RIILP, University of Wolverhampton, England

Ismail El Maarouf
Adarga Ltd., England

Michael Oakes
RIILP, University of Wolverhampton, England

This chapter is set in the context of Corpus Pattern Analysis (CPA), a technique developed by Patrick Hanks to map meaning onto word patterns found in corpora. The main output of CPA is the Pattern Dictionary of English Verbs (PDEV), currently describing patterns for over 1,600 verbs, many of which are acknowledged to be multiword expressions (MWEs) such as phrasal verbs or idioms. PDEV entries are manually produced by lexicographers, based on the analysis of a substantial sample of concordance lines from the corpus, so the construction of the resource is very time-consuming. The motivation for the work presented in this chapter is to speed up the discovery of these word patterns, using methods which can be transferred to other languages. This chapter explores the benefits of a detailed contrastive analysis of MWEs found in English and French corpora with a view on English-French translation. The comparative analysis is conducted through a case study of the pair (*bite*, *mordre*), to illustrate both CPA and the application of statistical measures for the automatic extraction of MWEs. The approach taken in this chapter takes its point of departure from the use of statistics developed initially by Church & Hanks (1989). Here we look at statistical measures which have not yet been tested for their ability to discover new collocates, but are useful for characterizing verbal MWEs already found. In particular we propose measures to characterize the mean span, rigidity, diversity, and idiomaticity of a given MWE.

Patrick Hanks, Ismail El Maarouf & Michael Oakes. 2018. Flexibility of multiword expressions and Corpus Pattern Analysis. In Manfred Sailer & Stella Markantonatou (eds.), *Multiword expressions: Insights from a multi-lingual perspective*, 93–119. Berlin: Language Science Press.
DOI:10.5281/zenodo.1182593

1 Introduction: phraseology and Multi-Word Expressions

Traditionally, people have long believed that each word has one or more meanings and that these meanings can be selected and put together, as if in a child's Lego set, to construct propositions, questions, etc. This belief is still widely (and unquestioningly, unthinkingly) held by many NLP (Natural Language Processing) researchers among others. This may indeed be a good way of accounting for basic *propositional logic*, but it accounts at best for only a very limited subset of natural language use. An alternative view is that logics are by-products of natural language. At the very least, we may say that the relationship between language and logic is not well understood. If the "Lego set" theory of meaning in language were tenable, it would not have been necessary for NLP and AI (Artificial Intelligence) researchers such as Ide & Wilks (2007), after many years of intensive (and expensive) effort, to declare that projects in Word Sense Disambiguation (WSD) have failed to achieve even their most basic goals.

> At present, WSD work is at a crossroads: systems have hit a reported ceiling of 70%+ accuracy (Kilgarriff et al. 2004), the source and kinds of sense inventories that should be used in WSD work is an issue of continued debate, and the usefulness of stand-alone WSD systems for current NLP applications is questionable. (Ide & Wilks 2007: 15).

The alternative view mentioned here is supported by lexicographers such as Atkins et al. (2001), Kilgarriff et al. (2004), and Hanks (2000). These lexicographers argue that much of the meaning of an utterance is carried by underlying patterns of co-selection of the words actually used, rather than by simple concatenations. These conclusions overlap to some extent with the tenets of Construction Grammar, though the methodologies are very different. In corpus linguistics, Sinclair declared, after a lifetime's empirical research into texts, corpora, and meaning, "Many if not most meanings require the presence of more than one word for their normal realisation" (Sinclair 1998: 4).

If these lexicographers and corpus linguists are right, it might appear that MWEs play a central role in the meaningful use of language. They are not merely an irritating set of exceptions, as used to be thought. According to this, MWEs are not exceptions to the rule; they are the rule. The exceptions, insofar as they exist in normal language use, are isolated meaningful uses of single words.

It has long been obvious that the meaning of MWEs such as *of course*, *a ballpark figure*, and *spill the beans* is not compositional. No courses, ball parks, or beans are invoked by someone deconstructing the meaning intended by a speaker

who uses these expressions. However, extended analysis of large volumes of data leads to the somewhat unwelcome conclusion that the concept of a MWE may also be flawed, being nothing more than an attempt to extend the "Lego-set'' theory to cover some so-called *fixed expressions* such as *spill the beans* and *kick the bucket*. Here, the choice of lexical items is fixed: one cannot talk meaningfully, except perhaps in jest, about **tipping over the haricots* or **booting the pail*. However, even in these very fixed MWEs, certain grammatical alternations, in particular verb inflections, are normal and unremarkable.

More to the point is the fact that many other expressions, that at first sight might be considered compositional, are associated with a limited phraseology. They do not vary freely, but employ selectional variations drawn from within a (usually quite small) lexical set. Such patterns are found for many expressions that intuitions alone might encourage us to classify as fixed. Corpus evidence shows that people not only grasp at straws, they also clutch at straws and even seize on straws. Moon (1998) observes that *shiver in one's shoes* (meaning 'to be afraid') may at first seem to be a fixed expression, but in fact corpus evidence shows that every lexical item in the expression allows a modicum of variation: people quake in their boots, shake in their sandals, and she even found a mention of policemen quaking in their size fourteens. (English policemen are supposed proverbially to have big feet.) The meaning of the idiom is the same in all cases; the cognitive values of the lexical items are so similar as to be virtually identical; and yet the actual words used to realize the expression can be different.

Conversely, when we examine the corpus evidence for an expression that might uncontroversially be classified as compositional, such as (1),

(1) *the wind was blowing from the north*

we find that the utterer of this unremarkable little sentence is in fact activating the meaning by drawing on a pattern containing a small but open-ended lexical set of items alternating with *wind*: *gale*, *blizzard*, *hurricane*, *typhoon*, *breeze*, *air*, not to mention adjectival subclassifications such as *a hot dry wind*, *a cold wind*, *strong winds*, *the fenland winds*, *a unidirectional wind*. To these can be added some much rarer lexical items such as *tempest*, *trades*, and *zephyr*. At the other end of the sentence forming the prototype or stereotype for this particular pattern, we find a very much larger set of expressions functioning as adverbials of direction: *from the north*, including *from the south*, *from the sea*, *over a cliff face*, *up the street*, *through a spider's web*, and so on.

These very conventional expressions are best classified as realizations of non-compositional patterns rather than as compositional concatenations for a variety

of reasons. A prominent one is that the pattern so identified is contrastive: it is a set of stereotypical phrases that contrast with other uses of the words. For example, this pattern (see example 2) contrasts with other patterns having different meanings formed with the same verb, such as *to blow a whistle* and *to blow up a bridge*.

Another reason for seeking to identify patterns of verb use is that, once a pattern is established in the language or in the mind of a speaker, it can be exploited metaphorically and in other ways. Some typical exploitations of this pattern of the verb *blow*, found in the British National Corpus, are shown in examples (2)-(6).

(2) *Dennis Healey* [*a politician*] *wobbles about according to which way the wind is blowing.*

(3) *The winds of neo-liberalism are blowing a gale through Prague.*

(4) *Faint liberal breezes had been blowing through the Vatican since the second Vatican Council.*

(5) *…the winds of change that have blown through the energy business.*

(6) *The winds of fate blew for Jean Morris, winner of Middlesbrough Council's Captain Cook Birthday Balloon Race.*

Metaphorical exploitations bring in additional evidence that a pattern has become established. In the previous examples, the meaning can only be understood in relation to the *the wind blows* (not, say, *blowing up a bridge*), but cannot be confused with it, as there is no wind blowing literally.

The aim of this short introduction to MWEs was to set the study of MWEs in the broad context of phraseology, and stress the obstacles in the way of linguistic description. In order to understand and process meaning in text, it is necessary first to compile inventories of patterns of language use, which can be used as benchmarks against which actual utterances can be compared. The following section presents Corpus Pattern Analysis, a method for deriving patterns from corpora.

2 The Corpus Pattern Analysis framework

Corpus Pattern Analysis (CPA) is a research procedure designed to create empirically well-founded resources for NLP applications by combining interactively human data analysis and machine learning. It is based on the Theory of Norms and Exploitations (TNE, Hanks & Pustejovsky 2004; Hanks & Pustejovsky 2005;

Hanks 2013). TNE in turn is a theory that owes much to the work of Pustejovsky on the Generative Lexicon (Pustejovsky 1995), to Wilks (1975)'s theory of preference semantics, to Sinclair's work on corpus analysis and collocations (Sinclair 1966; 1987; 1991; 2004), to the Cobuild project in lexical computing (Sinclair 1987), and to the Hector project (Atkins 1992; Hanks 1994). CPA is also influenced by frame semantics (Fillmore & Atkins 1992). It is complementary to FrameNet. Where FrameNet offers an in-depth analysis of semantic frames, CPA offers a systematic analysis of the patterns of meaning and use of each verb. Each CPA pattern can in principle be plugged into a FrameNet semantic frame. Some work in American linguistics (Jackendoff 2002) has complained about the excessive "syntactocentrism" of American linguistics in the 20th century. TNE offers a lexicocentric approach, with opportunities for synthesis, which will go some way towards redressing the balance.

CPA starts from the observation that whereas most words are very ambiguous, most patterns have one and only one sense. Each word is associated with a number of patterns based on valency, which is comparatively stable, and one or more sets of preferred collocations, which are highly variable (Hanks 2012). In CPA, patterns of word use are associated with statements of meaning, called IMPLICATURES. Each pattern has a primary implicature (the meaning of the pattern), and possibly a number of secondary implicatures (de Schryver 2010). To take a simple example, the word *blow* is multiply ambiguous. However, the expression *blow your nose* is unambiguous and contrasts with 60 or 70 other patterns of use of the same verb.

In the Pattern Dictionary of English Verbs (PDEV; http://pdev.org.uk), the main output of CPA, the sense of *blow your nose* is stored in the pattern "[[Human]] blow {nose}" while in the sense of "the wind blows" is represented by the pattern "[[Wind | Vapour | Dust]] blow [No object] [Adverbial of direction]". Patterns may combine various kinds of categories such as semantic types (Human, Wind, Vapour, Dust), grammatical categories (Adverbial of direction) and lexical items (*nose*). Semantic types are taken from the corpus-driven CPA Semantic Ontology available at http://pdev.org.uk/#onto. These categories may fill slots in the pattern template based on the SPOCA model, an acronym standing for the main clause roles that may be filled by arguments of a verb in a proposition: a Subject, a Predicator, an Object, a Complement, and an Adverbial (Halliday 1994). Each argument can in turn be further characterized if the pattern requires it, by filling information on the "subargumental cues" such as the nature of determiners, modifiers, quantifiers, prepositional phrases, and adverbs or particles.

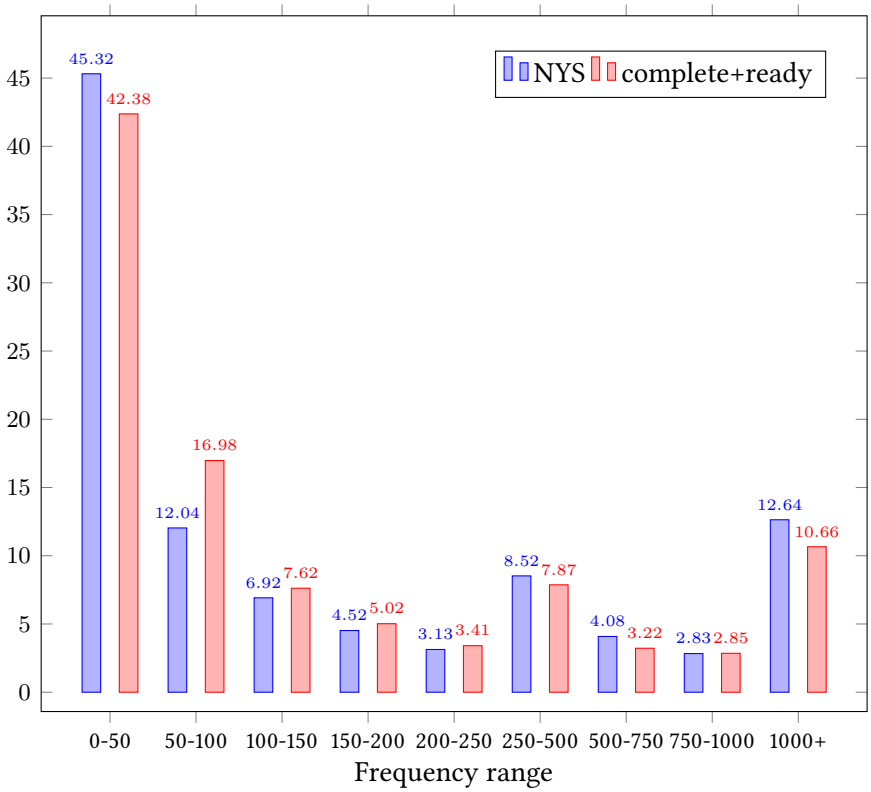

Figure 1: Proportion of NYS and complete and ready verbs w.r.t. frequency range in BNC50.

At the time of writing, PDEV covered 1,614 verbs for a total of 6,163 patterns, out of an estimated 5,500 total number of verbs in English (PDEV is therefore about 30% complete). PDEV is linked to a portion of the British National Corpus (BNC), BNC50, from which some of the statistics presented in this chapter are computed. BNC50 contains about 54 million tokens, and BNC, about 100 million. Figure 1 shows that the frequency distribution of complete verbs is very similar to that of NYS (Not Yet Started) verbs, e.g. that 40 to 45% of English verbs have a frequency lower than 50 in BNC50. For this reason, although PDEV is incomplete, it contains a representative sample of English verbs, large enough to warrant pilot studies. Results will need to be confirmed when PDEV is complete.

In PDEV, most verbs have a low number of patterns: the average number of patterns per verb is 3.8, and the verb with the greatest number of patterns is *break*, with 83 patterns. More than a quarter of verbs have only one pattern and

78% of verbs have five patterns or less. Verbs can also be contrasted in terms of qualitative characteristics. Particularly, some of them are used in idioms, others as phrasal verbs, and others combine with other lexical items in set phrases, that we propose to call lexically grounded patterns. Table 1 indicates the number of entries and patterns for these MWE-related categories of verbs.

Table 1: Number of verbs and of patterns for several MWE categories in PDEV.

MWE type	**# verbs**	**# patterns**	**% patterns**
Lexically grounded patterns	458	1,126	18.3
Phrasal verb patterns	198	512	8.3
Idiom patterns	200	453	7.3
MWE total	**548**	**1,649**	**26.7**

A *lexically grounded pattern* is a pattern which takes a lexical item or lexical set as an argument, either in subject, object, complement or adverbial position. For instance "[[Human]] take {responsibility} for [[Anything]]" is an example where a lexical item, *responsibility*, occupies the object position. In general the presence of lexical items is a strong sign of fixedness, so a significant portion of lexically grounded patterns overlap with idioms. All in all, there are 1,649 MWE patterns in PDEV, which accounts for 26.7% of PDEV patterns (about 34% of verbs). As each pattern is linked to a set of examples from the BNC, the whole MWE pattern set is connected to a total of 26,392 corpus examples (an estimated 84,836 over the whole BNC50, i.e. 1,545 per million).

PDEV *idioms* show very diverse statistical properties internally. For instance the estimated frequency in BNC ranges from 1 (e.g. *blowing off steam*) to 1,071 occurrences (for *as follows*) in BNC50, with an average frequency of 23.5 examples in BNC50 and a high standard deviation (67.2). 70% of idioms have 5 or more associated examples and 90% have less than 40 examples. The verb with the highest number of idioms is *throw*, with 24 idioms. Verbs with idioms on average have, for 64% of them, one idiom, for 19% of them, two idioms, and for 17%, three or more idioms.

3 A CPA study for English-French translation

The case study presented in this section focuses on *bite*, because it was found to encapsulate a large number of facts about English verbs, and particularly id-

iomatic structures. This verb is compared to the French *mordre*, which translates to 'bite' in its primary literal meaning: 'using teeth to cut'. We will observe how these verbs are used in each language, identify their common features and divergences by applying CPA to corpora. *Bite* was analysed using a sample of 500 lines from the BNC, and the same sample size was extracted from the Frtenten corpus (11 billion words; Jakubíček et al. 2013) for *mordre*.

Bite and *mordre* share interesting similarities in terms of their syntactic and semantic properties. Both verbs are mostly direct transitive, see examples (7) and (8), and can sometimes be accompanied with a locative adverbial, to indicate the [[Body Part]] bitten. Both verbs are also used in an intransitive pattern where the bitten entity, typically found in object position, is moved to a prepositional complement position, with *into* (*dans* in French) as preposition, see examples (9) and (10).

(7) *Those dogs bit the neighbours, the dustbin men, visiting aunts and each other.*

(8) *Le propriétaire ou le détenteur d'un chien qui a mordu une personne ou un autre animal a l'obligation de le déclarer au commissariat de son arrondissement.*

'The owner or the holder of the dog which has bitten a person or another animal is under the obligation of declaring it to the district police.'

(9) I'll wager that your salivary glands started pumping out liquid as you imagined yourself biting into the lemon.

(10) *Je mords dans une pêche : un goût d'eau sucrée accompagné d'un sentiment de vide.*

'I bite in a peach: a taste of sweet water together with a feeling of emptiness.'

These syntactic patterns are frequently employed in different situations which sometimes share very little in common with the literal meaning of the verb. To contrast these uses, CPA entries make use of semantic types which characterise the semantic properties shared by the collocates found in a given syntactic slot. In the literal sense of transitive patterns, *bite* and *mordre* typically collocate with [[Human]] (with the particular case of vampires) and [[Animal]] (e.g. *dogs*) as subjects, and with [[Human]], [[Animal]], and [[Body Part]] as objects. Other [[Physical Object]] nouns (e.g. *pillows*, *coins*, *pencils*) were found in English, but not in French, although they could be found in a larger sample. Transitive

patterns of *bite* were also found to combine with [[Eventuality]] as subject and [[Human]] or [[Institution]] as objects, as in example (11).

(11) *Provincial had been bitten by its own success.*

In this case, the pattern means "[[Eventuality]] adversely affect [[Human]] or [[Institution]]". The construction *bite + into* was also found with a metaphorical pattern expressed as "[[Event]] bites into [[Event]]", sharing the same meaning as the previous pattern (signaling an adverse effect). These metaphorical uses of *bite* seem to be English-specific: no such pattern was found in the French sample. This is because French typically uses *ronger* 'gnaw', as in example (12).

(12) a. *The recession is biting deeply into industry.*
b. *La recession ronge durement l'industrie.*

When English speakers use *bite* with direct objects such as *nails* or *fingers* to mean 'chewing at one's fingernails, biting the tips off', French speakers use *ronger* for *ongles* and *doigts* respectively. In this case, it is also considered as a distinct pattern in French. Other patterns were found, such as "[[Physical Object 1]] bite in|into [[Physical Object 2]]",[1] where the subject is neither [[Human]] nor [[Animal]]. This pattern can only be translated to French with *mordre* to cover uses where "[[Blade]] makes small cuts into [[Physical object]]". When the subject is *acid*, signalling the corroding effect the acid has on metal, French uses *ronger*. For other types of object nouns, such as *ploughs*, French would use the phrasal expression *se planter + dans*.

Semantic types can also help to contrast existing patterns from uses which combine with specific animals, e.g. [[Snake]], which was found both in French and English, and which refer to a different situation, defined as "[[Snake]] stabs [[Human]] or [[Animal]] with fangs, typically injecting poison under the skin". However, when considering [[Insect]] (e.g. *mosquitoes*) in subject position, the normal French verb is *piquer* (see example (13) below).[2]

(13) a. *The mosquitos came up and bit me in the dark.*
b. *Les moustiques sont venus et m'ont piqué dans le noir.*

However, *bite* does not collocate with nouns of other flying bugs such as *wasps*, *bees*, or *hornets*,[3] whereas these nouns can be used indifferently with *piquer*. This language-specific feature can be explained by the extra-linguistic fact that insects bite to feed, but bees, wasps, and hornets possess a specific device, positioned at

[1] English also uses patterns with the phrasal verb *eat away* for this meaning.
[2] Although *mordu par les moustiques* is acceptable.
[3] English uses the verb *sting*.

the bottom of their bodies, used to kill or in self-defence. This is the only pattern where *piquer* can be used as a translation of *bite*. The pattern "[[Human]] or [[Animal]] bite through [[Physical_Object]]" also has a literal meaning, but cannot be translated using *mordre*. The best translation equivalent appears to be *grignoter* (literally *nibble*), since it keeps the notion of 'using teeth', and correctly translates 'insects biting through leaves'. However this verb does not translate the fact that the bitten entity is filled with holes. The verb *bite* was only found in a single intransitive use, "[[Process]] bites", with the meaning "[have] a noticeable effect, usually an adverse effect", as in *the recession bit deeper*. This would be translated into French with the expression *se faire sentir* (literally 'to be felt'). The verb *mordre* was also found in metaphorical patterns which could not be translated with *bite*, namely "[[Building]] mord [[Area]]", as in (14), and "[[Vehicle]] mord la route", as in (15).

(14) *Certaines des constructions mordaient sur des terres privées.*
'Some of the buildings **encroached** on private lands.'

(15) *Quand vient le temps d'effectuer un dépassement, le véhicule mord la route.*
'When the time comes to pass the car in front, the vehicle **grips** the road.'

In addition to these patterns, 6 idioms were found for *mordre* (see Table 2), and 10 idioms for *bite* (see Table 3).

Table 2: Idiom CPA patterns for the verb *mordre*.

No	Pattern / *Implicature*	Frequency	%
4	[[[Human]] | le poisson] mord ([à l'hamecon | à l'appat]) [[*Human*]] *takes the bait (= is lured to do something that has bad consequences)*	10	2
7	[[Human]] mord {la vie à pleines dents} [[*Human*]] *enjoys life to the full* [*literally, *bites life with full teeth*]	6	1.2
9	[[Human]] se mord {les doigts} [[*Human*]] *experiences a bitter time* [*literally, *bites his/her fingers*]	21	4.2
11	[[Human 1]] fait mordre [la poussière] [à [[Human]]] [[*Human 1*]] *causes* [[*Human 2*]] *to bite the dust (= to die) or to lose a challenge* [*the latter sense only in French*]	6	1.2
12	[le serpent] se mord [la queue] [[*Human*]] *is stuck in a* [[*State of affairs*]] *and cannot find a way out* [*literally, *the snake bites his own tail*]	16	3.2
16	[[Human]] ne mord pas [NO OBJ]] [[*Human*]] *does not bite (= is harmless)*	6	1.2

Table 3: Idiom CPA patterns for the verb *bite*.

No	Pattern / *Implicature*	Frequency	%
13	Human 1 bites Human 2's head off *Human 1 speaks sharply and unkindly to Human 2*	5	1.22
14	Human bites REFLDET lip *Human grips his or her lip firmly with the teeth*	8	1.96
15	Human bites off more than [[Human]] can chew *Human undertakes a task that is too difficult for him or her to accomplish successfully*	4	0.98
16	Human bites the hand that feeds [[Human]] *Human attacks his or her benefactor*	5	1.22
17	Human or Institution bites the bullet *Human or Institution decides to do something necessary but unpleasant*	21	5.13
18	Human is bitten by the [MOD] bug *Human becomes very interested in [MOD]*	7	1.71
19	Human bites the dust *Human dies suddenly and violently*	2	0.49
20	Entity or Process bites the dust *Entity or Process comes to a sudden and unwelcome end*	8	1.96
21	Human bites REFLDET tongue *Human makes a desperate effort not to say what is on his or her mind*	8	1.96
22	Once bitten twice shy *An unpleasant experience causes someone to be more cautious in future*	3	0.73

These idioms share little in common (apart from the correspondence between patterns 11 in French and 19 in English) and do not involve the notion of 'teeth cutting'. Thus the correct French to English translation (and vice versa) required knowledge that is encoded in CPA patterns. Pattern 12, for instance, *le serpent se mord la queue*, is used to refer to situations where *serpents* 'snakes' are not involved, a phenomenon generally referred to as non-compositionality. In the next section we propose to measure this property as well as other important features, such as rigidity, using statistical measures. These measures will be applied to idioms which will be the focus of §4.

4 Statistical measures for the characterisation of MWEs

In this section, we will describe the use of statistical measures to automatically characterise the flexibility of MWEs. We feel that this is an important research

topic, as it can contribute to describing in which respects MWEs are flexible and help to speed up their extraction from corpora.

4.1 Word association measures and lexicography: PMI

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 & 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 (Point-wise Mutual Information) between word x and word y is given by the formula

(16) $$I(x, y) = log_2 P(x, y)/P(x).P(y)$$

where P(x,y) is the probability of the two words occurring in a common context (such as a window of 5 consecutive words), while $P(x)$ and $P(y)$ are the probabilities of finding words x and y respectively anywhere in the corpus. PMI 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 & Hanks 1989). PMI was used by Church & 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 Act]] and [[Physical Object]], giving at least two senses of the verb (Hanks 2012).

4.2 Span, rigidity, diversity and idiomaticity

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*. We propose to characterise the flexibility of a multiword expression using four statistical measures, each focusing on a dimension of variation.

A MWE can be characterised by its mean span MEAN SPAN, that is, the stretch of text it is found to cover on average. This can be measured using the mean μ of the relative distances between two words making up the MWE, and computed as follows:

(17) $$\mu\,(X, Y) = \frac{1}{n}\sum_{i=1}^{n} dist(X_i, Y_i)$$

A MWE can also be further characterised by its RIGIDITY. This can be measured using the standard deviation σ of the relative distances between the two words:

(18)
$$\sigma(X,Y) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(dist(X_i,Y_i) - \mu(X,Y)\right)^2}$$

For standard deviation, the minimum value when all the examples have identical span is 0, and there is no theoretical upper limit. Higher values 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 span, 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:

(19)
$$E = -\sum_{i=1}^{N} p_i \log_2 p_i$$

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 MWE 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.

Finding statistical evidence for the flexibility of a sequence of words does not automatically entail that all the examples of the sequence belong to a MWE, and

that the reading is non-literal. We therefore propose to measure the IDIOMATICITY of a MWE in context, by taking the ratio of the number of idiomatic occurrences of the expression divided by its total number of occurrences:

(20)
$$Idiomaticity\,(x,y) = \frac{number\ of\ idiomatic\ occurrences}{total\ number\ of\ occurrences}$$

A value of 1 would indicate that the MWE is always idiomatic, while a lower value would indicate that the MWE can be ambiguous with respect to its idiomatic reading. It must be borne in mind that this equation depends on a number of factors such as the overall frequency of the verb of a MWE in the specific language. The more frequent in everyday language the constituents of the MWE are, the more probable for them to be encountered in a corpus in their literal meaning. This is not related to the idiomaticity of the expression per se, which has to do with the opacity of the expression: the more opaque (as opposed to transparent) it is, the more idiomatic it is.

4.3 Worked out example

To illustrate how the values of each measure are computed, we propose a worked out example based on a pair of words used as boundaries, *bite* and *dog*, in a sample of 10 examples taken from pattern 1 of *bite*:

(21) "[Human 1 | Animal 1] bites [Animal 2 | Physical Object | Human 2]"

We chose this pair because it is a strong collocation (PMI = 7.7 in BNC). To apply our statistical measures, the first thing to do is to compute the distance between the boundary words. First, it is worth noting that we lump together alternative surface forms of the same boundary word, so we consider both *dogs* and *dog* as one word. Different decisions at this stage may lead to different results.

Figure 2 provides an example using signed distance (left or right): in the first example, *bite* is four words away to the left of *bite*, the distance is therefore -4. To compute the mean span, however, we recommend using the unsigned distance (i.e., 4 for the first example), but it is important to use the signed distance to compute the standard deviation, in order to capture word order variation. The unsigned text distances are therefore, in order of appearance of the examples, 4,4,3,2,4,1,3,2,1,2.

The mean μ characterises the mean span of an expression: *bite* and *dog* are 2.6 words apart.

(22)
$$\mu_{\{bite,dog\}} = \frac{4+4+3+2+4+1+3+2+1+2}{10} = 2.6$$

The standard deviation characterises the rigidity of an expression and makes use of the mean of the signed distances μ' computed as follows:

(23)
$$\mu'_{\{bite,dog\}} = \frac{(-4)+(-4)+3+(-2)+4+(-1)+3+(-2)+(-1)+(-2)}{10} = -0.6$$

The standard deviation can therefore be computed as:

(24)
$$\mu'_{\{bite,dog\}} = \sqrt{\frac{(-4-(-0.6))^2+(-4-(-0.6))^2++(-2-(-0.6))^2}{10}} = \sqrt{\frac{76.4}{10}} = 2.76$$

The score obtained for *bite* and *dog* is indicative of a low rigidity (2.76).

To compute diversity (entropy), we extract all the patterns of word forms between boundaries and count the frequency of each pattern class. Again, characters could also be used as the basic unit, but we use words; the string of the pattern can be characterised in various ways, we use word forms. A pattern is a full string between boundaries, with the null class accounting for cases where boundary words are adjacent. For X = {*dog*,*dogs*}, Y = {*bite*,*bites*,*bit*,*bitten*}, and i = {*that barks doesn't*, *that had been*, *another. In*, *to*, *by a police*, { }, *his pet*, *are*, *always*},

'A dog^{-4} that^{-3} barks^{-2} doesn't^{-1}	bite^{0}	1	,' replied Antonio Navarro,
of the dogs^{-3} that^{-2} had been^{-1}	bitten^{0}	2	and strayed: scared that th
in saliva when one animal	bites^{0}	3	another^{1}. in^{2} dogs^{3} , one of t
who had trained his dog^{-2} to^{-1}	bite^{0}	4	Arabs, and who informed
/p> He was chased and	bitten^{0}	5	by^{1} a^{2} police^{3} dog^{4} and then a
t was saved when her dog^{-1}	bit^{0}	6	him. The 22-year-
heltenham yesterday after	biting^{0}	7	his^{1} pet^{2} dog^{3} , which was at
time by their own dog^{-2} are^{-1}	bitten^{0}	8	in the bedroom. In our bree
d. After that dogs^{-1}	bit^{0}	9	me on the feet. Blood came
/ herself that dogs^{-2} always^{-1}	bite^{0}	10	people, especially them. T

Figure 2: Example of calculated distances for the pair (*bite*, *dog*) in concordance for *bite*.

P_i corresponds to the number of times the string is observed in the sample, divided by the total number of examples (in our case, 10). The entropy is computed as follows:

(25) $E_{\{bite,dog\}} = -\left(\left(\frac{1}{10}\log_2\frac{1}{10}\right) + \left(\frac{1}{10}\log_2\frac{1}{10}\right) + + \left(\frac{1}{10}\log_2\frac{1}{10}\right)\right) = 3.12$

The entropy is quite high as there is no particular pattern that dominates the sample: only the null pattern occurs twice, but the others, only once. Finally, no expression formed with *bite* and *dog* was found to have an idiomatic reading, therefore the idiomaticity is equal to 0.

The proposed measures are described for two variables. However, many idioms include more than two words, such as *let the cat out of the bag*. In such cases we take the span of the idiom as the distance from the first word to the last, which for this example would be 6 words.

5 A contrastive statistical analysis of idioms

In a pilot experiment on the annotated sample of the BNC corpus of *bite*, 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 span 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 spans was relatively small (0.48), reflecting that in all cases but one the variation consisted of the insertion of a single word, but the diversity index was at its maximum value for a set of 6 examples, log2(6) = 2.585.

The results for *bite* were borne out when the experiment was repeated on a larger corpus, the entire BNC. Table 4 shows the results obtained for English idioms. Idioms are represented by their boundary words and the table provides the scores using standard measures of collocational strength (PMI, t-score, and Log-Dice), along with the absolute frequency and our new measures: idiomaticity, entropy, mean span, and standard deviation.

In the full BNC, there were 19 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. 16 of these were idiomatic, including 3 variants of the farewell *sleep tight, don't let the bed bugs bite*, and 3 were literal as in *I've been bitten by bugs in a hooker's bed*. This gave an idiomaticity of 16 / 19 = 0.842. Of the idiomatic examples, almost all were unique, such as *bitten by the travel bug* - the other *bugs* included *puppy love*, *acting*, *racing*,

Table 4: Summary of scores for some idioms of *bite*. SD: Standard Deviation

Idiom	Freq (total)	PMI	t-score	Log-Dice	Idiomaticity	Entropy	Mean span	SD
[back] [bite]	87	5.914	10.380	5.549	0.989	0.338	1.057	0.277
[bullet] [bite]	36	10.484	6.477	8.561	1	1.069	2.055	0.404
[head] [bite] [off]	30	6.009	7.639	5.600	0.775	3.281	3.032	2.721
[dust] [bite]	26	8.918	5.088	7.438	1	0.235	2.03	0.192
[bug] [bite]	19	10.589	4.688	7.894	0.842	3.326	3.125	2.578

flower pressing and *showbiz*. 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 pattern classes of 4, 3 and 3 members and 9 unique occurrences, had a very high value of 3.326. In terms of rigidity, the mean distance between *bite* and *bug* was 3.125, with a high standard deviation of 2.578. This was because cases such as *the acting bug really bit me* used the inchoative alternation, so *bug* appeared before *bit*. Also, influencing rigidity was the fact that even in the active voice, the number of intervening words[4] could vary.

The MWE *bite the bullet* occurred in 36 sentences altogether, there were no literal examples at all, but that MWE appeared as mentions of both a racehorse and a pop song. Of the other 34 examples, the vast majority (29) were exactly in the form "[bite] the bullet", the remainder being in the forms *bit the ideological bullet* (3), reversed as in a *harder bullet to bite* (1), and a statement by President Bush about an opponent: "I bite bullets, he bites nails". The idiom was rather rigid, with a mean span of 2.055, and a fairly low standard deviation of 0.404. Diversity was also low at 1.069.

The results on French idioms were obtained from the tagged sample from the Frtenten corpus. The results obtained here have taken only a part of the corpus into account. In the future, we will perform an exhaustive analysis of the remaining 196,500 examples. The scores are given in Table 5, using the same headers as Table 4.

The idiom *le poisson mord à l' hameçon* is a popular expression in French which means 'to take the bait' (see Table 2). As illustrated in Table 5, it was found in 3 different forms, which, despite varying mean span and frequency, were each

[4]Its French adjectival counterpart, *mordu de* is also diverse: *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.'

Table 5: Summary of scores for some idioms of *mordre* (500 lines sample). SD: Standard Deviation

Idiom	Freq (total)	PMI	t-score	Log-Dice	Idiomaticity	Entropy	Mean span	SD
[poisson] [mordre]	4	7.340	1.988	-2.222	0.75	0	1	0
[hameçon] [mordre]	4	12.259	2	2.664	0.75	0	3	0
[appât] [mordre]	2	10.475	1.413	0.894	1	0	3	0
[vie] [mordre]	6	4.434	2.523	-5.126	1	1.792	2.833	0.372
[doigt] [mordre]	21	9.387	4.789	-0.174	0.952	1.08	1.190	0.154
[poussière] [mordre]	6	9.360	2.446	-0.204	1	0	1	0
[queue] [mordre]	16	9.670	4.118	0.109	1	0.34	1.187	0.527
[serpent] [mordre]	13	11.022	3.604	1.457	1	1.7	1.846	0.591

found to be maximally fixed (standard deviation = 0) and minimally diverse (entropy = 0).

The pattern "[[Human]] se mord {les doigts}" rarely took its literal meaning in French, standing for 'a person experiencing a bitter time for his past actions' in 20 cases out of 21. It usually occurred in the corpus as *mordre les doigts*, but sometimes as *se 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 of 1.19, a standard deviation of 0.15, and an entropy of 1.08.

The idiom *mordre dans la vie à pleine dents* was also found as *mordre la vie à pleines dents*. Table 5 lists the scores when both variants are combined. If we consider *vie* as the boundary word (*à pleines dents* was only found once in a mention of a song), *mordre dans la vie* occurred 4 times with mean span of 3, while *mordre la vie* was found twice with mean span of 2.5. Since 5 out of 6 examples had a distance of 3 words, the standard deviation was quite low (0.372); however the idiom had a high entropy (2.833), as *mordre la vie* contributed 2 different unique pattern classes to the idiom.

If we compare English and French, the corresponding phrases *mordre la poussière* and *bite the dust* both have standard deviations for their spans of 0, since in the BNC and Frtenten corpora the verb is always exactly 2 words before the noun. However, as can be seen in Table 2, *mordre la poussière* can have the additional use 'losing a challenge' which was not found in English. The MWEs *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. It is

worth noting that translation systems unaware of these facts 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 span, Shannon Diversity, and idiomaticity give reasonable results as they reflect the flexibility of a MWE. We could also suggest a measure of constructional flexibility, which might be the ratio of times a MWE occurs in the active voice divided by the number of times the MWE occurs altogether, whether in the active or passive voice.

6 Generalization of statistical measures

Evaluating the applicability of statistical measures to different languages is one way to evaluate their validity. This section describes other methods to test the generalizability of measures.

6.1 Comparison with cognitively salient idioms

Hanks (2013: 5,21,214) makes a distinction between expressions that are cognitively salient (roughly equivalent to "easily called to mind") and socially salient (roughly equivalent to "frequently used"). He suggests that cognitive salience and social salience are independent variables, or may even be in an inverse relationship: that is, frequently used expressions are buried deep in the language user's subconscious mind and are not necessarily easily called to mind. The idioms *kick the bucket* and *spill the beans* are probably the most cognitively salient and most frequent idioms cited by linguists. Other idioms cited in this chapter are *grasping at straws*, *the way the wind blows*, or *shivering in one's boots*. These idioms, along with 4 idioms involving *bite*, make up the set of 10 idioms used for the experiments described in this section.

In the BNC, *kick the bucket* has 21 occurrences, although another 4 sentences containing both words were discounted as *kick* and *bucket* appeared in separate clauses. Another 8 were from a linguistic discussion of the phrase, as in "notice 'kick the bucket' appears as a verb phrase". Only 5 were idiomatic, in the sense of *to die*: 4 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 3.5 words, a standard deviation of 1.870, and a modest diversity of 0.721. 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 5/21 = 0.238.

In contrast, the phrase *spill the beans*, found 42 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: 40/42 = 0.952. 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 mean separation was 2.025, the rigidity as measured by the standard deviation was 0.987, and diversity as measured by entropy was a lowish value of 0.370. 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 reports that MWEs like *spill the beans* are more flexible than the relatively well behaved *kick the bucket.* Although *kick the bucket* is more idiomatic in the sense that it is fully opaque, it occurs more often in the text in its literal meaning because its literal meaning is more frequent in everyday language.

An idiom which stands out in Table 6 is *way the wind blows*, which was by far the strongest collocation according to the t-score and LogDice measures, and the lowest idiomaticity score (or having the greatest proportion of literally-intended examples). *bite …bug* had highest entropy, as one can metaphorically be bitten by many kinds of bug. Finally *bite …hand …*[*benefit*] had the greatest mean span and standard deviation of span.

6.2 Inter-annotator agreement

Another way of demonstrating the validity of a statistical measure, such as MWE idiomaticity or mean span, is to determine the Inter-Annotator Agreement (or Inter-Rater Reliability, IRR). This is the degree to which two or more observers might concur on a classification or annotation task. A measure is only valid to the extent that humans can agree on the classification of the individual instances

Table 6: 10 English idioms retained for generalization experiments.
SD: Standard Deviation

Idiom	Freq (total)	PMI	t-score	Log-Dice	Idioma-ticity	Entropy	Mean span	SD
[back][bite]	87	5.914	10.380	5.549	0.989	0.338	1.057	0.277
[bullet][bite]	36	10.484	6.477	8.561	1	1.069	2.055	0.404
[head][bite] [off]	30	6.009	7.639	5.600	0.775	3.281	3.032	2.721
[bug][bite]	19	10.589	4.688	7.894	0.842	3.326	3.125	2.578
[hand][bite] [BENEFIT]	15	5.584	7.639	5.196	1	2.463	5.933	5.842
[bean][spill]	40	10.947	6.705	8.917	0.952	0.370	2.025	0.987
[straw] [grasp/clutch]	33	9.865	6.077	8.172	0.892	2.213	3.485	1.623
[way][wind] [blows]	21	10.663	25.264	10.652	0.676	2.488	3.5	0.534
[shoe/boot] [quake/shiver /shake]	12	5.043	5.056	5.608	1	2.057	3.417	1.382
[bucket][kick]	5	8.647	4.349	7.004	0.238	0.721	3.500	1.870

which contribute to the measure. For example, do they agree on whether a MWE is being used in its idiomatic sense or not, and where it starts and ends? IRR falls in the range 0 for only random agreement to 1 for perfect agreement. As an illustration, we estimated the IRR, using Krippendorff's α measure,[5] between two native speakers of English as regards the span and idiomaticity of the phrase *kick the bucket*. There were 26 sentences in the British National Corpus containing both *kick* and *bucket*. An α value of 1 denotes perfect agreement among the annotators, and 0 shows that agreement occurred only by chance. The instructions given to each annotator were as follows:

> For each sentence, choose one of the following:
>
> 1) the phrase *kick the bucket* (or a grammatical variant of it) does not appear in the sentence;
>
> 2) the phrase *kick the bucket* (or a grammatical variant of it) is idiomatic, and means 'to die';

[5] Krippendorff's α may be calculated using the 'irr' package in the R statistical programming language. The package 'irr' can be installed by the following command: `install.packages("irr", repos = "http://cran.r-project.org")`

The annotators' responses should be stored in a matrix, where each row corresponds to annotators' response values. The R command to create a matrix for three examples and 2 annotators is, for example: `m = matrix(c(1,1,3,3,1,2), nrow=2)`.

> 3) the phrase *kick the bucket* (or a grammatical variant of it) is literal, and actually means to physically kick a bucket.
>
> If you answered 2) or 3), use [to show where the phrase *kick the bucket* begins and] to show where it ends, as in the example: "I'm too young to [kick the bucket]".

In our experiment to find the agreement of the native speakers as to whether the phrase *kick the bucket* was absent, literal or idiomatic, Krippendorff's α was 0.745.[6] A value between 0.6 and 0.8 is said to be "good" agreement (Altmann 1991: 404).

This experiment was modified to consider only those cases where the annotators considered the phrase *kick the bucket* to be present:[7] we were looking at the agreement between the annotators in distinguishing literal and idiomatic uses, and Krippendorff's α was 0.635, still "good".

To look at the agreement with respect to the span of the idiom, the values in the matrix were replaced with the number of words between the square brackets marked by the annotators, or NA if they did not find the idiom in the sentence.[8] The annotators agreed in every case where they both marked off the start and end of the idiom (α = 1), showing that the limits of the idiom *kick the bucket* were clear-cut to these native speakers. Thus according to this small experiment, the measures of idiomaticity and mean span are valid for the expression *kick the bucket*.

6.3 Correlation and relatedness of measures

While the previous section illustrated techniques to test the validity of statistical measures, this section describes a final experiment focusing on the relatedness of different measures. To do this, we compared the values of our set of 10 idioms (see Table 6) according to 10 measures. These included the measures of mean and standard deviation of idiom span, Shannon Diversity and idiomaticity, compared with four standard measures for collocation strength: frequency of collocation, PMI, t-score and LogDice. Both the t-score and LogDice are used by the Sketch Engine lexicographers' tool (Rychlỳ 2008).

[6] Krippendorff's α is found by the following command: `irr:kripp.alpha(m,"nominal")`.

[7] The matrix was modified so that all the 1s (denoting absence of the phrase) were replaced by "NA" (not applicable).

[8] This type of numeric data is called "ratio" data, so the appropriate command to calculate Krippendorff's α is: irr:kripp.alpha(m,"ratio").

To determine whether these measures were independent of each other or whether one acts as a predictor for another, Spearman's rank correlation coefficient was computed for each pair of measures. This statistic was preferred to the Pearson correlation coefficient, as the sets of values for some of the individual measures were not normally distributed. The correlations between the measures are shown in Table 7. The most statistically significant correlation (p = 0.002, cor = 0.88) was between PMI and LogDice, suggesting that these measures of collocational strength agree with each other well. Another significant correlation was the inverse correlation between frequency and mean span (p = 0.008, cor = -0.78). Thus there was a tendency for more frequent idioms to be shorter (and to a lesser extent, not statistically significant) more rigid in their structures. There was no significant correlation between any of the measures in Table 4 and Table 5 with either of the measures of collocational strength, frequency and PMI.

Table 7: Correlations between scores in the 10 idiom study.
SD: Standard Deviation

Idiom	Freq (total)	PMI	t-score	Log-Dice	Idiom-aticity	En-tropy	Mean span	SD
Freq	1							
PMI	0.36	1						
t-score	0.51	0.03	1					
LogDice	0.24	**0.88**	-0.04	1				
Idiomaticity	0.23	-0.42	-0.09	-0.30	1			
Entropy	-0.39	0.08	0.01	0.01	-0.29	1		
Mean span	**-0.78**	-0.19	-0.15	-0.07	-0.25	0.43	1	
Standard deviation	-0.61	-0.27	-0.30	-0.44	-0.20	0.62	0.55	1

These results suggest that the new measures of idiomaticity, entropy, mean span and standard deviation of span may not be useful for discovering new MWE, but as we have shown, are useful for describing the characteristics of MWE once discovered.

7 Conclusions and perspectives

Sinclair (2004) wrote that the so-called "fixed phrases" are not in fact fixed: most phrases in English display some variety of form. "Variation gives the phrase its essential flexibility, so that it can fit into its surrounding context". Conversely,

each word cannot be considered as a simple "Lego brick" which can be fitted in a slot-and-filler system, as corpus-based investigations reveal that each word preferentially selects other words, echoing J.R. Firth's maxim that "You should know a word by the company it keeps". In this context we have proposed to use Corpus Pattern Analysis as a technique to describe word patterns found in corpora, and have applied this technique to two verbs in French and English. CPA is a corpus-based technique to detect the lexical, syntactic, and semantic preferences of verbs, such as the fact that *bite* preferentially selects *mosquitoes* and *bugs* while *sting* normally selects *bees*, *wasps* and *hornets*. The application of the CPA methodology to a French corpus revealed however that *mordre*, the French translation of *bite* in examples such as *dogs bite*, was neither used with *mosquitoes* nor with *bees*: French speakers prefer to use *piquer* 'sting' for most kinds of "flying entity aggression". This suggested that patterns of words are more reliable units of translation than words in isolation, which opens up new research perspectives for using CPA in Translation studies and Machine Translation.

In this chapter, we proposed to use statistical measures which could be applied to any MWE in any language, by illustration on French and English. These new statistical measures characterise the flexibility of a MWE based on text distance: the mean span of MWEs, the standard deviation of the distance between their boundary words, their internal diversity, and their idiomaticity ratio. The results obtained by the application of these measures to *bite* and *mordre* revealed that each captured useful features of MWEs which compared favourably with intuitive notions of flexibility and compositionality. It is worth noting that the implementation of these measures required us to make a number of decisions explicitly, particularly deciding on a basic unit such as the word or character. Perspectives include testing these measures on other languages, particularly those with so-called free word order, and application to Machine Translation.

In his analysis of extended units of meaning, Sinclair (1991) noted, as we have done in our discussion of *bite* and *mordre*, that idioms can carry across to other languages. In his example, the Italian equivalent of *naked eye* is *a occhio nudo*. While this is true of many expressions, the contrastive analysis proposed in this chapter also suggests that the semantic space occupied by a single lexical item can be covered by several lexical items in another language. The MWE *naked eye* also exhibits a phenomenon we have not examined in this chapter: there is greater consistency of patterning to the left of the collocation than to the right. This suggests that we could use our measures to find the rigidity or diversity not only of the MWE itself, but of its context on either side. We could also look for the semantic prosody associated with MWE – for example, things seen with the naked eye tend to be difficult ("small", "weak" or "faint") to see.

Abbreviations

AI	Artificial Intelligence	PDEV	Pattern Dictionary of English Verbs
CPA	Corpus Pattern Analysis	PMI	Point-wise Mutual Information
IRR	inter-rater reliability	WSD	Word Sense Disambiguation

References

Altmann, Douglas G. 1991. *Practical statistics for medical research*. London: Chapman & Hall.

Atkins, Beryl T. 1992. Tools for computer-aided corpus lexicography: the Hector project. In Ferenc Kiefer, Gábor Kiss & Julia Pajsz (eds.), *Papers in Computational Lexicography: COMPLEX'92*, vol. 115, 1–60. Budapest: Hungarian Academy of Sciences.

Atkins, Beryl T., Núria Bel, Pierrette Bouillon, Thatsanee Charoenporn, Dafydd Gibbon, Ralph Grishman, Chu-Ren Huan, Asanee Kawtrakul, Nancy Ide, Hae-Yun Lee, Paul J. K. Li, Jock McNaught, Jan Odijk, Martha Palmer, Valeria Quochi, Ruth Reeves, Dipti Misra Sharma, Virach Sornlertlamvanich, Takenobu Tokunaga, Gregor Thurmair, Marta Villegas, Antonio Zampolli & Elizabeth Zeiton. 2001. *Standards and Best Practice for Multiligual Computational Lexicons. MILE (the Multilingual ISLE Lexical Entry) Deliverable D2.2-D3.2*. ISLE project: ISLE Computational Lexicon Working Group. http://www.w3.org/2001/sw/BestPractices/WNET/ISLE_D2.2-D3.2.pdf, accessed 2018-4-19.

Church, Kenneth Ward & Patrick Hanks. 1989. Word association norms, mutual information and lexicography. In *Proceedings of the 27th Association for Computational Linguistics Conference (ACL)*, 78–83.

de Schryver, Gilles-Maurice. 2010. Getting to the bottom of how language works. In *A way with words: Recent advances in lexical theory and analysis: a Festschrift for Patrick Hanks*, 3–34. Menha Publishers.

Fillmore, Charles J. & Beryl T. Atkins. 1992. Towards a frame-based organization of the lexicon: the semantics of RISK and its neighbors. In Adrienne Lehrer & Eva Feder Kittay (eds.), *Frames, fields, and contrasts: New essays in semantics and lexical organization*, 75–102. Hillsdale, NJ: Lawrence Erlbaum Associates.

Hanks, Patrick. 1994. Linguistic norms and pragmatic exploitations or, why lexicographers need Prototype Theory, and vice versa. In Ferenc Kiefer, Gabor Kiss & Julia Pajzs (eds.), *Papers in computational lexicography: COMPLEX'94*, vol. 94, 89–113. Hungarian Academy of Sciences.

Hanks, Patrick. 2000. Do word meanings exist? *Computers and the Humanities* 34(1). 205–215.

Hanks, Patrick. 2012. How people use words to make meanings: Semantic types meet valencies. In *Input, process and product: Developments in teaching and language corpora*, 54–69. Brno (CZ): Masaryk University Press Masaryk.

Hanks, Patrick. 2013. *Lexical analysis: Norms and exploitations*. MIT Press.

Hanks, Patrick & James Pustejovsky. 2004. Common sense about word meaning: Sense in context. In *Proceedings of text, Speech and Dialogue (TSD)*, 15–17. Brno, Czech Republic.

Hanks, Patrick & James Pustejovsky. 2005. A pattern dictionary for Natural Language Processing. *Revue Française de linguistique appliquée* 10(2). 63–82.

Ide, Nancy & Yorick Wilks. 2007. Making sense about sense. In Eneko Agirre & Philip Edmonds (eds.), *Algorithms and applications*, 47–73. Dordrecht, The Netherlands: Springer.

Jackendoff, Ray. 2002. *Foundations of language*. New York: Oxford University Press.

Jakubíček, Milos, Adam Kilgarriff, Vojtěch Kovvář, Pavel Rychlý & Vit Suchomel. 2013. The TenTen corpus fmily. In *7th International Conference on Corpus Linguistics (CL 2013)*. Lancaster.

Kilgarriff, Adam, Pavel Rychlý, Pavel Smrz & David Tugwell. 2004. The Sketch Engine. In *Proceedings of Euralex 2004*, 105–116. Lorient, France,

Moon, Rosamund. 1998. *Fixed expressions and idioms in English: A corpus-based approach*. Oxford University Press.

Oakes, Michael P. 2012. *Describing a translational corpus*. Michael Oakes & Meng Ji (eds.). Amsterdam/Philadelphia: John Benjamins. 115–148.

Pustejovsky, James. 1995. *The generative lexicon*. MIT press.

Rychlỳ, Pavel. 2008. A lexicographer-friendly association score. In *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN*, vol. 2008, 6–9.

Sinclair, John. 1966. Beginning the study of lexis. In Charles Ernest Bazell, John Cunnison Catford, Michael Alexander K. Halliday & Robert H. Robin (eds.), *In memory of J. R. Firth*, 410–430. London: Longman.

Sinclair, John. 1987. *The Collins Cobuild English Language Dictionary*. London: HarperCollins.

Sinclair, John. 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Sinclair, John. 1998. The lexical item. In Edda Weigand (ed.), *Contrastive lexical semantics*, vol. 171 (Current Issues in Linguistic Theory), 1–24. John Benjamins.

Sinclair, John. 2004. The search for units of meaning. In John Sinclair & Ronald Carter (eds.), *Trust the text: language, corpus and discourse*, 24–48. London: HarperCollins.

Smadja, Frank. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics* 19(1). 143–177.

Wilks, Yorick. 1975. A preferential, pattern-seeking, semantics for natural language inference. *Artificial Intelligence* 6(1). 53–74.