Patrick Hanks

Research Institute for Information and Language Processing, University of Wolverhampton

Bristol Centre for Linguistics, University of the West of England

How people use words to make meanings: Semantic types meet valencies

This paper proposes that meanings in text are both created and understood by matching actual text occurrences (or creations) against patterns of usage stored in the brain. A ‘pattern’ in this sense has two elements: valency, which is comparatively stable, and one or more sets of preferred collocations, which are highly variable. To understand collocations, we draw on prototype theory developed by the cognitive scientist Eleanor Rosch (1973a, 1973b), its philosophical counterpart developed by the philosopher Hilary Putnam (1970, 1975a, 1975b), and the linguistic insights of John Sinclair (1966, 1987, 1991, 2004). Collocates are grouped into lexical sets according to their semantic type, using the Generative Lexicon theory of James Pustejovsky (1995).

Corpus pattern analysis shows that each word habitually participates in only a comparatively small number of patterns, and that most patterns are unambiguous in their interpretation. This yields a new theory of language use – a ‘double helix theory’ called the theory of norms and exploitations (Hanks in press). This argues that language use is governed by not one but two interactive sets of rules: a set of rules for using words normally and a set of rules for exploiting the norms creatively.

1. Introduction

It is a truism that the meaning of a word is (to a greater or lesser extent) dependent on the context in which it is used. But what is ‘context’? And how is a relevant context to be recognized and distinguished from what Firth (1957: 187) called “the mush of general goings-on” in language? To understand something that is said or written, we need to map it in some way onto an underlying pattern of meaningful usage. The present paper is about phraseological patterns and how to discover them.

The essential first step is to get the valency right – the number of arguments that a word requires to enable it to be used correctly. Valency depends in part on the part of speech. Attributive adjectives typically have a valency of one: they are governed by a head noun, and that’s it. Verbs and predicative adjectives typically have a valency of between one and three: some combination of subject, direct object, and indirect object or adverbial. Valencies of nouns vary between zero and three, depending on the subcategorization of the noun. Valency theory was first developed by Lucien Tesnière in 1959, and provides a more practical basis for lexical and semantic analysis than generative grammar. Also useful for the analytic apparatus that we need is the slot-and-filler grammar of Michael Halliday (1961).

Let us invent an example, for the sake of exposition, showing some of the many things that a pattern is not:

1. *Matilda saw an ant sitting on a peacock.*[1]

In 1, the verb *saw* has a valency of two (subject and direct object) and the participle *sitting* also has a valency of two (subject and adverbial, the latter being realized as a prepositional phrase). This absurd invented sentence is *not* a pattern. It does not illustrate a pattern; it does not instantiate a pattern; it has nothing whatsoever to do with patterns. It is not even an exploitation of a pattern, as we shall see. It is, however, syntactically perfectly well formed, for the verb *see* regularly governs (among other things) –*ing* forms: ‘an ant sitting on a peacock’ instantiates an –*ing* form. Likewise, the verb *sit* regularly governs prepositional phrases headed by *on*. And so on. Moreover, in 1 the (supposed) selectional restrictions are respected: the verb *see* selects an event or a state of affairs as its direct object; ‘an ant sitting on a peacock’ represents (or rather, purports to represent) a state of affairs; the prepositional verb *sit on* selects a noun denoting a physical object as a prepositional object after *on*, and *peacock* indisputably denotes a physical object. We cannot object that the selectional restriction of *sit on* requires an inanimate prepositional object, for it is perfectly normal for people to sit on horses and camels, which are animate. Other animates are, however, less usual in this prepositional object slot. (So already here, the seeds of a problem with the notion of ‘selectional restrictions’ can be seen. I shall return to this point shortly.)

This tedious exposition of a silly invented sentence (of a kind beloved by speculative linguists) contains the seeds of both an unsatisfactory linguistic theory (one that is widely accepted) and a rather better one, which has not yet been fully elaborated. Having discussed the absurdity of a

[1] In this paper, the convention is followed of printing citations from actual usage (including examples taken from large corpora) in roman, while invented examples (usually invented for some contrastive purpose) are printed in italics.

sentence invented in the time-honoured manner of speculative linguistics, let us now turn and look at what words actually do. Language teachers nowadays generally encourage their students to ask, '*Do* you say this in English', rather than '*Can*

you?' Viable teaching practice and a viable linguistic theory must be based on how people actually use words, not on how they might possibly use words. But still, within this context, we must allow for occasional anomalous uses that are intentionally, purposefully creative.

At the heart of the unsatisfactory theory lies the notion of 'selectional restrictions'. This is attractive to theoretical and computational linguists, no doubt because a 'restriction' has predictive power. However, wishing that something were so does not make it so. After over two decades of work in corpus linguistics (e.g. Stubbs 2001; Wray 2002; Hanks 2004; Hoey 2005), the selection of lexical items in a valency slot is governed by a system of preferences, not a system of restrictions. Careless talk equating preferences with restrictions merely confuses the issue. If progress is to be made on the real predictive power of patterns, it is necessary to develop a system of probabilistic predictions based on selectional preferences. This is already being done in some versions of statistical language processing and probabilistic linguistics (e.g. Bod et al. 2003), and it is central to the work of Hanks (1994, in press) on linguistic norms and pragmatic exploitations. This theory is called the Theory of Norms and Exploitations (TNE).

At the heart of TNE lies the notion that all linguistic categories are based on stereotypes or prototypes and that analysis of meaning must therefore be statistically based. It thus makes predictions about *probable* usage and *probable* meanings, and refrains from speculating about the boundaries of possible usage and meaning. Current linguistic theories of usage and meaning regularly attempt – and regularly fail – to account for all possible uses of a linguistic item. TNE suggests that there is a principled reason for this failure, namely that, rather than a clear-cut boundary between well-formed and ill-formed sentences, there is instead a vast grey area of possible but increasingly unlikely exploitations of normal usage.

Stereotypes and prototypes admit analogies and calculations of probabilities – how *probable* is it that anything will be said to *sit on* an animate entity? And of course, the answer springs to mind: very probable, if the animate is a horse or a camel; less probable, otherwise. If the animate is a bird, it is no doubt possible, but it is certainly not normal. Introspection of this kind can be confirmed and extended by examination of patterns in corpora. Compounding the linguistic (pragmatic) misdemeanour of our well-formed but bizarre invented sentence is the fact that, normally, ants don't sit – not on peacocks, nor an anything else. But in terms of word use, this, too, is a probability, not a certainty. It may be a physiological impossibility for ants to sit, but that does not prevent an inventive linguist (or anyone else) creating a syntactically well-formed sentence in which an ant is said to be sitting on something.

It seems almost too obvious to say that language teaching and language learning need to be based on a clear understanding of how words are actually used, rather than on speculation about imagined possibilities. When second-language learners use a word in the second language creatively, it is important that both they and the readers or hearers understand that this is what they are doing, and that they are not merely making a mistake.

To illustrate how people really do use words normally and meaningfully, I will first present a detailed corpus-driven examination of the uses of the word *shower*, not as a noun, but as a verb. Like so many English verbs, *shower*, when used to denote an action or an event, may be regarded as a grammatical metaphor. Prototypically, *shower* is a noun. Underlying the word, at a sub-wordclass level, is some perceptual or experiential notion of water falling rapidly in droplets from above – from the clouds above, or from a purpose-built artefact in a bathroom or perhaps beside a swimming pool. This fundamental cognitive prototype is vague, not precise. It does not even have a part of speech. But if we turn now and ask about patterns of usage, the part of speech becomes very important, as do the valencies and collocates. Let us see how this works in the case of this verb.

2. Sometimes, valency can differentiate meanings

Occasionally, valency alone is sufficient to make a semantic distinction. Thus two senses of the verb *shower* (broadly, 'wash the body under flowing water' and 'donate in large quantities') are distinguished by the number of arguments. Sentence 2 below is intransitive (i.e. it has a valency of one), governing neither a direct object nor a prepositional object. By contrast, 3 and 4 have a valency of three: both these examples are transitive, with an adverbial (prepositional phrase).

2. He showered and dressed quickly.
3. He showered her with kisses.
4. *He showered kisses on her*.

There is very little difference in meaning between 3 and 4; at most, there is a difference of focus. The underlying valency patterns of 2-4 are:

A. [NP] shower.
B. [NP1] shower [NP2] {with [NP3]}.
C. [NP1] shower [NP3] {on [NP2]}.

Comparing the frequencies in a randomly selected sample of 100 sentences from the British National Corpus (BNC), we find that only 16 of them have the 'wash' sense (valency pattern A). Multivalent uses are much more common, though less literal.

There is also a 2-valent intransitive (inchoative) use, as in 5:

5. Bits of broken glass showered over me.

D. [NP1] shower [NO OBJ] Adv/[NP2].

Theoretically (at least), we can also predict a transitive 2-valent usage:

E. [NP1] shower [NP2].

But this is rare. I did not find any examples of pattern E in the 100 million words of the BNC. However, we can imagine examples such as 6:

6. *She showers the dog every Sunday*.

The phrase 'every Sunday' in 6 is an optional adjunct, not an argument of the verb. It has no effect on the meaning of the verb. So the invented – possible but less likely – example 6 illustrates *shower* with a valency of two. For this verb, then, the main factor determining the meaning is the presence or absence of an adverbial argument (typically realized in the form of a prepositional phrase: 'with [NP]' or 'on [NP]').

The transitive 2-valent usage of this verb is an excellent example of a possible but rare pattern, posing a modest dilemma for the lexical analyst. Although this pattern does not occur in the BNC, and although there are serious theoretical objections to speculating about possible patterns in the absence of evidence showing actual usage in earnest, it would not be unjustifiable to add this pattern to a pattern dictionary. A Google search does turn up a smattering of genuine transitive 2-valent uses, such as example 7 from an Australian forum page for trainee nurses:

7. She will also have to shower and clean male patients.

The dilemma for the analyst is whether to classify this rare 2-valent transitive usage as a 'normal' pattern or as an exploitation of a pattern. The dilemma can be resolved, partly by giving frequency information, and partly by appealing to the level of generalization. In a systematic pattern dictionary such as the Pattern Dictionary of English Verbs (see Hanks & Pustejovsky 2005), statements of comparative frequency are given. So, for example, in a sample of 100 uses of the verb *shower* in the BNC, intransitive (1-valent) uses account for 16% of the sample, while trivalent uses (*shower* NP1 *with* NP2, and *shower* NP2 *on* NPt) account for 69%. The inchoative pattern D, 'something showered on somewhere', accounts for 13%. I venture to suggest that these percentages will remain stable over any reasonably large sample of any general corpus of British, American, or Australian English. The 2-valent transitive use exemplified in 6 and 7 could, then, be classified as a pattern, but it will still account for less than 1% of samples. Alternatively, a transitive use like this could be classified as a syntactic exploitation of valency pattern A.

For most uses of a word, difference of valency alone is not necessarily a meaning determinant. For example, the verb *accept* is usually 2-valent: it has a subject and a direct object. However, in certain contexts, the direct object can be omitted without affecting the meaning, as in 8:

8. *He invited her to dinner and she accepted*.

This use, with an actualized valency of only one, has the same meaning as the normal use with valency two. It exploits the norm.

3. Generally, collocations and semantic types are also needed

So far, so good. Valencies are capable of making some semantic distinctions and, as we shall see, they are an essential component of many others. However, in most cases of semantic ambiguity, valency analysis alone is not enough. Even increasing the delicacy of valencies by introducing thematic roles such as Agent, Patient, Beneficiary, and Instrument does not get us very much further. Something different in kind is needed. In a word, that something is *collocation*. All content words prefer the company of one or more lexical sets of other words and, thanks to developments in corpus linguistics, these preferences can now be analysed. The nature of lexical sets is discussed by Hanks and Jezek (2008) and Jezek and Hanks (2010). Below are some sentences from the BNC in random order:

9. Boris showered the woman with presents.
10. Rather than the hoped-for cash, they were showered with snuff boxes and other trinkets, to Leopold's disgust.
11. Lauren Bacall, Bianca Jagger… and Lionel Blair were among the stars who showered him with praise.
12. If they ignore the remark or reply negatively they may be accused of rudeness and/or showered with abuse.
13. You long to shower gifts on everyone.
14. European heads of government… showered telegrams of congratulation on Clinton.
15. Despite all the criticisms showered on this model during the past forty years, it still occupies the center of the stage.
16. Chinese parents do, of course, shower love and attention on their children.
17. Whistling and swearing offends them and they will shower the guilty person with pebbles and gravel until he stops.
18. The eruption showered debris on Pompeii.
19. The DC-10 exploded, showering them with debris.

How many of the words in these sentences are statistically significant collocates of the verb *shower*? To answer this question, we would need to consult a much larger sample and see how many of these words recur, and how often. Moreover, we need to take account of the comparative frequency or rarity of each collocate in the corpus as a whole. A rare word such as *plaudits* is much more significant (*because* it is rare) if it recurs several times in relation to the verb *shower* than a common word such as *love*. Church and Hanks (1990) showed how this could be done computationally using a large corpus. Since that date, many other techniques for measuring statistically significant collocations have been developed, for example in the SketchEngine software of Kilgarriff et al. (2004). Different statistical measures give different results. Different results are suitable for different applications. For example, t-score tends to favour collocates that are high frequency function words, so it is suitable for applications such as identifying prepositional choice in particular contexts. Pointwise mutual information, the statistic used by Church and Hanks (1990), favours rare content words, so it is more suitable for analysing meaning through collocation.

Once a set of statistically significant collocates has been identified, they can be sorted into different lexical sets. A lexical set is a group of words that share one or more semantic features. This semantic feature is very often the word's 'formal', to use Pustejovsky's (1995) term, i.e. a superordinate concept.

The collocates found in the prepositional object slot of the verb phrase '*shower* [someone] *with* [something]' includes at least the following (recurrent) lexical items found in large corpora:

> *abuse, accolades, affection, applause, arrows, attention, awards, cash, compliments, debris, dollars, favo(u)rs, flowers, gifts, glass, hono(u)rs, jewel(le)ry, kisses, largesse, love, money, obscenities, plaudits, praise, presents, rocks, shrapnel*

At first, this may seem rather a confusing jumble of words. However, as we read through it, the natural human instinct to sort and classify kicks in. We start to see shadowy outlines of patterns. Moreover, once the 'seed members' of one or more lexical sets have been identified, the sets can be augmented with other, less frequent collocates, on the grounds of shared semantic properties. And the resultant patterns almost always turn out to be semantically contrastive. That is, each pattern (identified by a combination of valency and collocates) has a distinctive meaning.

A salient pattern involves *showering someone with plural* [[Speech Act]]s[2]. These are either very positive (*accolades, compliments, plaudits, praise*) or very negative (*abuse, obscenities*). Closely associated with these are words denoting attitudes, mostly positive, and not necessarily expressed in speech (*attention, applause, favours, honours, affection, kisses, love*). Next, we note that you *shower someone with* physical objects that are pleasant to receive (*gifts, presents, awards, jewellery, flowers, cash, dollars, money*). In all the cases mentioned so far, the prepositional object of *with* typically correlates with a grammatical subject denoting a human agent.

20. Lauren Bacall, Bianca Jagger... and Lionel Blair were among the stars who showered him with praise.
21. if they ignore the remark or reply negatively they may be accused of rudeness and/or showered with abuse.
22. Boris showered the woman with presents.
23. Rather than the hoped-for cash, they were showered with snuff boxes and other trinkets, to Leopold's disgust.

Examples 20 and 22 have the verb in the active voice: the subject is explicitly present (*Lauren Bacall, Bianca Jagger, and Lionel Blair; Boris*). These words all have the semantic type [[Human]], expressing an intrinsic semantic property shared by all these nouns and many thousands of others which can also occupy the same slot in relation to the verb *shower*. *Shower* in this sense is a transitive verb, and active transitive verbs regularly alternate with a passive construction, as in 21 and 23. Here, of course, the underlying subject or Agent has not been made explicit. The reader is constrained (by pattern matching in his or her head) to assume that the Agent is [[Human]].

[2] Semantic types are conventionally written in double square brackets and stored in a hierarchical ontology. For additional information, see Pustejovsky et al. (2004).

Contrasting with these is a small but different set of words in the prepositional object slot denoting inanimate physical objects which it is distinctly unpleasant to be on the receiving end of: *rocks, pebbles, gravel, debris, arrows, glass,* and *shrapnel*. Here, the correlating subject is not restricted to [[Human]]; there is also a typical (but *not* a necessary) correlation with grammatical subjects denoting events such as an explosion or the eruption of a volcano:

24. whistling and swearing offends them and they will shower the guilty person with pebbles and gravel until he stops.
25. The eruption showered debris on Pompeii.
26. The DC-10 exploded, showering them with debris.

4. A syntactic alternation

So far, we have noted a fairly narrow range of prepositional objects activating a closely related range of set of meanings of the verb *shower*, and we have noted that these meanings are activated regardless of whether the verb is active or passive. This is absolutely standard for transitive verbs: it is unusual for passive sentences to activate a different meaning of a verb from active sentences. The main purpose of the passive is to obviate the necessity of stating an agent explicitly, not to activate a different meaning.

There are many other alternations of verb patterns. Different verbs participate in different alternations. Syntactic alternations generally involve a difference of emphasis rather than a difference of meaning. In the case of *shower*, there is an alternation which shifts the prepositional object into the direct object slot while shifting the direct object into a prepositional object slot governed by *on*. This may be classed as a syntactic alternation, rather than a different pattern, because no difference in meaning is activated.

27. You long to shower gifts on everyone.
28. Despite all the criticisms showered on this model during the past forty years, it still occupies the center of the stage.
29. Chinese parents do, of course, shower love and attention on their children.
30. The eruption showered debris on Pompeii.

It is clear from these examples that all 3-valent senses of the verb *shower* participate in this alternation. There is no sense distinction of *shower* that depends on the syntactic prepositional *with/on* distinction.

In the current version of the Corpus Pattern Analysis project, each of these alternations is treated as a separate pattern. This has the effect of doubling and in some cases even quadrupling the number of patterns for every verb that participates in such alternations. It would probably make more sense, therefore, to treat regular alternations of this kind as subsets of patterns rather than as patterns in their own right. For this to work effectively, empirical evidence that each such alternation is actually used must be adduced.

5. Grouping lexical items into sets

Every argument of every pattern of every verb is realized by a lexical set of nouns (which may include multi-word items such as *snuff boxes*). A lexical set may consist of anything from a single word to a vast array of lexical items. On the basis of what has been said so far, it will be clear that a crucial question for effective language processing is: can the contrasting lexical items found in the various arguments (or 'valency slots') in relation to each verb be arranged into groups according to some common semantic property?

- SET 1 (Semantic type [[Physical Object]]): Typical set members: *gifts*, *presents*, *jewellery*...
  Less typical set members: *trinkets*, *snuff boxes*.
- SET 2 (Semantic type [[Speech Act]]): Typical set members: *praise*, *abuse*, *insults*.

It would be nice if these two lexical sets were sufficient to justify a distinction into two different patterns, F and G:

F. [[Human1]] shower [Human2]] {with [[Physical Object]].
G. [[Human1]] shower [Human2]] {with [[Speech Act]]}.

However, as we have seen, that generalization would be inadequate to capture the true nature of these two patterns. Several additional observations are needed. In the first place, the prepositional object in both patterns must be either a mass noun or a plural noun. This looks suspiciously like a necessary condition. It is not possible to *shower* someone with *an expensive watch* or with *a word of advice*. Secondly, the semantic type [[Physical Object]] is not sufficient to make rather an important semantic distinction: not all Physical Objects are Gifts. The distinction between *showering* someone with *pebbles* and showering them with *snuff boxes* has implications for the intentions of the subject of the verb, namely *giving* vs. *attacking*.

Typically (but not necessarily), collocates can be grouped into sets according to their shared semantic type, but it turns out to be quite difficult to decide on the appropriate level of generalization for a semantic type. For this reason, it is helpful to make a distinction between the intrinsic semantic properties of a concept and the extrinsic properties assigned to a word by the context in which is used.

If the verb *shower* has an inanimate subject (physical or abstract), this normally correlates with a [[Physical Object]] as second argument and [[Human]] or [[Location]] as third argument: *explosions shower debris on people and places* (or *shower people and places with debris*). Correlations of this sort among the semantic arguments predict the meaning of the verb, which can be discovered by a procedure that Church and Hanks (1990: 28-29) call 'triangulation':

> Despite the fact that a concordance is indexed by a single word, often lexicographers actually use a second word such as *from* or an equally common semantic concept such as a time adverbial to decide how to categorize concordance lines. In other words, they use two words to *triangulate in* on a word sense. This triangulation approach clusters concordance lines together into word senses based primarily on usage (distributional evidence), as opposed to intuitive notions of meaning. Thus, the question of what is a word sense can be addressed with syntactic methods (symbol pushing), and need not address semantics (interpretation), even though the inventory of tags may appear to have semantic values.
>
> The triangulation approach requires "art." How does the lexicographer decide which potential cut points are "interesting" and which are merely due to chance? The proposed association ratio score provides a practical and objective measure that is often a fairly good approximation to the "art." Since the proposed measure is objective, it can be applied in a systematic way over a large body of material, steadily improving consistency and productivity.

6. Exploitation of normal patterns

The story so far – mapping lexical sets onto valencies – goes a very long way towards explaining how people use words to make meanings. The apparatus we have outlined here accounts for the way in which a very substantial proportion of ordinary language utterances create meanings. But before we finish, we must take account of two additional phenomena: exploitations of norms and contextual roles.

Exploitations of normal patterns have been mentioned several times already in this paper. Any linguistic regularity may be exploited for rhetorical, comic, or other effect. As we saw in example 8 above, the normal syntactic structures in which a word is used may be exploited by ellipsis – but also in other (increasingly convoluted) ways. For example, it is well known that in English normal word order, which is normally used to assign clause roles such as subject, object, and prepositional object, may be exploited under certain circumstances for emphasis, as in 31:

31. Now this, I don't approve of.

A very common type of exploitation involves an anomalous argument. There is nothing remotely difficult about the interpretation of 32, so it may not be immediately obvious that this is an exploitation of a norm. However, if we ask, 'What is the normal semantic type of subjects of the verb *punish*?' we will find that it is not normally a procedure such as *rehabilitation*. This, then, is an exploitation by reason of the anomalous argument:

32. Whatever the intention, *rehabilitation* does *punish* people; in particular, it allows people to be put into institutions where they would rather not be.
(BNC) Bob Roshier, 1989. *Controlling Crime: The classical perspective in criminology*.

Anomalous arguments are a rich source of creative figurative language, which depends on a wide variety of exploitation rules. The full typology of exploitation rules is too extensive to discuss in detail here; a fuller discussion will be found in Chapter 8 of Hanks (in press). What creative figurative expressions have in common is that they involve unusual and unexpected collocations, activating various kinds of cognitive resonance. Thus the meaning of 33 is perfectly clear, but nevertheless it is unusual to talk about bricks ‘arriving’. The writer evidently had some metaphorical or rhetorical purpose in mind in choosing this particular phraseology.

33. As I sat down to write up my diary a brick *arrived* through my sitting room window.
(BNC) M. Grist, 1993. *Life at the Tip: Les Bence on the game*.

It is important to distinguish between conventional metaphors, which are no more than second-order regularities, and creative metaphors, which exploit norms.

7. Semantic types vs. contextual roles

The theoretical foundation for semantic types lies in the Generative Lexicon theory (GL) of Pustejovsky (1995). According to GL, content words draw on four different kinds of resource for their meaning:

i. **Event Structure** specifies the event type of the clause or expression, for example Action, Process, State;
ii. **Argument Structure** specifies the number of arguments of a predicate;
iii. **Lexical-Type Structure** defines the semantic type of a word in a hierarchical ontology of concepts, for example [[Human]], [[Artefact]], [[Vehicle]];
iv. **Qualia Structure** provides a basis for structural differentiation of the predicative force of a lexical item.

If we ask, what is the ‘event type’ involving the verb *shower* with one argument, the answer is [[Activity]]. Activities in this sense are actions that humans do (contrasted with Processes). Only humans *shower*. Although it is undoubtedly possible that any creature or indeed any physical object may be caught in a *rain shower* (or, for that matter, placed under the *shower* in a bathroom), *showering* as a verb implies intentionality and bathrooms. Logical possibility and regularities of word use are quite different things.

In passing, likewise, we may note that although, logically, given the normal syntagmatic behaviour of weather terms in English, one ought to be able to predict an expression with a dummy subject, *It was showering*, equivalent to *It was raining*, there is no evidence for the existence of such an expression.

Now, on the other hand, if we ask, what is the event type involving the verb *shower* with three arguments, the answer will be, ‘Well, that depends...’ In particular, it depends on the semantic type of the third argument – [[Speech Act]] or [[Physical Object]], as the case may be.

However, even this apparatus (valency and collocations classified according to semantic type) is sometimes insufficient for unambiguous analysis. To complete the analysis, one more thing is needed, namely an identification of the contextual roles assigned by the pattern as a whole. This is because many aspects of meaning are activated by the phraseology in which words are used, not merely by the intrinsic semantic properties of the words in themselves.

Contextual roles are semantic properties assigned by the context. Consider 34:

34. *Mr Woods sentenced Bailey to seven years.*

Here, the semantic type of both *Mr Woods* and *Bailey* is [[Human]]. It should be noted that this is a high probability, not a certainty. In isolation, the expressions *Mr Woods* and *Bailey* almost certainly denote human beings, but it is possible, for example, that someone has a dog called *Mr Woods* or *Bailey*.

The point at issue is that the roles [[=Judge]] and [[= (Criminal) Offender]] are assigned by the context, not intrinsic semantic properties of the nouns used. The same is true of the contextual roles [[= Punishment, = Imprisonment]] assigned to the expression *seven years* in this context.

Applying this to the verb *shower*, we may note that *shower* takes many words denoting Physical Objects as arguments, but the question whether the physical objects in question are missiles or gifts is a matter of contextual role, not intrinsic semantics of the nouns used.

8. Collocational analysis of nouns

Let us turn now to the corpus-driven analysis of the meaning of nouns. This, too, can proceed by examining and organizing collocates, though the procedure and results are different in kind from collocational analysis of verbs[3] as the combinatorial properties of nouns are not as strong as those of verbs and adjectives. Very often, relevant collocates are found somewhere in the general environment of a target noun, rather than in a strict syntagmatic relationship with it. Thus occurrence of the word *nurse* anywhere near *doctor* provides a good reason for selecting a medical sense rather than, say, the sense ‘holder of an advanced academic qualification’ for *doctor*. The *doctor* and the *nurse* do not have to be in a syntagmatic relationship for this selectivity to succeed linguistically.

Take, for example, the noun *spider*. Here, valency is not always relevant: analysis of several large corpora shows that *scorpion* and *cockroach* are among the most salient collocates of *spider*. As collocates, they do not regularly stand in any particular syntactic relation to *spider*, however; they are simply found nearby, in a window of five words to the left or right of the node word *spider*.

Collocates, selected from corpora, can be used as a basis for building up a ‘cognitive profile’ that consists of phraseologically well-formed, idiomatic statements for the noun (see Figure 1). The cognitive profile starts with a general classificatory statement that is not corpus-derived, but the remaining statements derive from collocational analysis. Relevant collocates are highlighted in italics.

[3] It would be more precise to say that the collocational analysis of nouns denoting entities is different from that of words denoting events and states of affairs. Some nouns (‘deverbal nouns’) denote events, and their analysis is necessarily verblike.

Figure 1. Corpus-driven cognitive profile of the noun *spider*

Spider: *a spider* is a **living creature**. Even big spiders are quite small compared with humans.

- Types of spider: many thousands of species of spiders are known (including *funnel-web, web-building, orb-weaving, bird-eating, ground-dwelling, giant, huge, large, tiny, poisonous, black widow, camel, redback, trapdoor, wolf, whitetail, crab* spiders.) And *tarantulas*.
- Some species of spiders *hunt prey*.
- Some spiders *bite*.
- Some species of spiders are *poisonous*.
- Many species of spiders *spin webs*, with threads of *strong silk*.
- Spiders *lurk* in the centre of their *webs*.
- Spiders *control* what is going on in their *webs*.
- Spiders have eight *legs*.
- Their legs are *thin*, hairy, and long in proportion to body size.
- Spiders have *eight eyes*.
- Spiders spend a lot of time being *motionless*.
- Spiders' *movement* is *sudden*.
- Spiders *crawl*.
- Spiders *scuttle*.
- Spiders are *swift* and *agile*.
- Spiders can *run up walls*.
- Many people have a *dread* of (*hate*) or are *frightened* of spiders.
- People *kill* spiders.
- In folk taxonomy, *scorpions* and *cockroaches* are often classified together with spiders as creepy-crawly creatures.

The goal of a noun cognitive profile such as this is to organize as many as possible of the salient collocates of the target word into meaningful, informative, and idiomatic statements. A good cognitive profile uses all the salient collocates of the target word and so provides excellent guidance on its idiomatic use. Eagle-eyed readers will notice a theory of semantic types lurking behind the classificatory statement 'animate entity'. Classification by semantic types (i.e. conceptual hypernyms or superordinate terms, as in WordNet: http://wordnet.princeton.edu/) is even more noticeable in Figure 2.

If we apply this kind of collocational analysis of corpus data to *shower* as a noun, we find that it has at least four cognitive profiles, as set out in Figure 2.

In the right circumstances, it can be a useful exercise for intermediate language learners to use corpus data and a statistical profiler such as SketchEngine to compile cognitive profiles like this for nouns whose meaning they understand at least partly or approximately, but with whose idiomatic usage they are not fully conversant.

Figure 2. Corpus-driven cognitive profile of the noun *shower*

Shower 1: *a shower* is a **weather event** of short duration.

- Typically, a shower is a short downpour of *rain*.
- Other types of showers: in cold weather there are *snow* showers, *wintry* showers, and showers of *hail and sleet*.
- A shower may be *heavy* or *light*.
- Weather forecasters talk about *scattered* showers, *occasional* showers, or *the odd shower*.
- Showers *sweep over* or *across* locations.
- After a short time, showers *die away* or *die out*; showers *clear*.
- People get *caught in* a shower.
- *April showers* are (supposedly) short and refreshing, in a period that is otherwise sunny.
- Metaphorically, physicists speak of showers of *particles*; astronomers speak of showers of *meteorites* or *meteors*.

Shower 2: *a shower* is an **artefact** for pouring water in droplets simulating rainfall.

- Typically, a shower is *provided* by an architect or house designer and *installed* by a builder, either in a *cabinet* in the *bathroom* of a house, or above the *bath*, or in a separate *shower-room*.
- An *en-suite* shower is one that is installed in a room adjacent to a *bedroom*.
- When installed correctly, a shower *works*.
- Types of shower: there are several trade names for different types of shower. Some showers are *electric* showers or *power* showers. (Others are gravity-fed.)
- People *switch* (or *turn*) a shower *on* in order to *use* it and *off* after using it.
- A shower is also a location with such an artefact fixed high up in it so that it can pour water from above, such that a person *stands in* the shower in order to *wash* his or her *hair* or *body*.

Shower 3: *a shower* can also be a **human activity**, using a shower (profile 2) to wash the whole body and the hair.

- A person *takes* a shower.
- A shower may be *hot, cool*, or *cold*.
- Taking a shower is *refreshing*.

Shower 4: In informal spoken English, a group of useless, unwanted human beings may be referred to as *a shower*.

9. Conclusion

The detailed corpus analyses in this paper have shown how collocations affect idiomatic language use, and how (for verbs or 'event words') valency interacts with collocations to determine meaning.

Shower is a fairly simple verb, but its behaviour in everyday usage, as recorded in the BNC, is sufficiently complex to illustrate some general principles that apply to the meaning analysis of all verbs:

- Valency makes a major contribution to the foundation of meaning of verb phrases, and, since the verb is the pivot of the clause, this implies a major contribution to the meaning of sentences.
- An equally important contribution is made by collocations.
    - Noun collocates are sorted into lexical sets according to their intrinsic semantic type. These play an additional important role in determining the meaning of a verb in context.
    - In some cases, context imposes a particular interpretation (a 'contextual role') on a lexical set. Contextual role is an extrinsic semantic property of the set, and should not be confused with the semantic type.
- Additional constraints that are sometimes relevant include number (singular, mass, plural), axiological (good/bad) evaluation ("semantic prosody" in Sinclair's 1991 terminology), and syntactic alternation.
- Any linguistic regularity may be exploited for some special effect. But exploitations are themselves rule-governed.

Corpus analysis of nouns ('noun-y' nouns) is both more straightforward and less satisfying. Verbs typically interact with other linguistic items more than nouns; nouns typically have more interaction with the (assumed) physical realities of the external world , and (partly for this reason) the linguistic patterns of noun use tend to be much less constrained than those of verbs.

References

Bod, R., J. Hay & S. Jannedy (eds). 2003. *Probabilistic Linguistics.* Cambridge MA: MIT Press.

Church, K.W. & P. Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics* 16(1): 22-29.

Firth, J. 1957. *Papers in Linguistics 1934-1951*. London: Oxford.

Halliday, M.A.K. 1961. Categories of the theory of grammar. *Word* 17(3): 241-292.

Hanks, P. 1994. Linguistic norms and pragmatic explanations, or why lexicographers need prototype theory and vice versa. In F. Kiefer, G. Kiss & J. Pajzs (eds), *Papers in Computational Lexicography: Complex '94*. Budapest: Research Institute for Linguistics, Hungarian Academy of Sciences.

Hanks, P. 2004. Corpus pattern analysis. In G. Williams & S. Vessier (eds), *Euralex Proceedings, Vol. 1*. Lorient: Université de Bretagne-Sud, p. 87-97.

Hanks, P. In press. *Lexical Analysis: Norms and exploitations*. Cambridge MA: MIT Press.

Hanks, P. & E. Jezek. 2008. Shimmering lexical sets. In E. Bernal & J. De Cesaris (eds), *Proceedings of the XIII EURALEX International Congress*. Barcelona: Universitet Pompeu Fabra, p. 391-402.

Hanks, P. & J. Pustejovsky. 2005. A pattern dictionary for natural language processing. *Revue Française de Linguistique Appliquée* 10(2): 63-82.

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

Jezek, E. & P. Hanks. 2010. What lexical sets tell us about conceptual categories. *Lexis* 4: 7-22. http://screcherche.univ-lyon3.fr/lexis/IMG/pdf/Lexis_4.pdf, accessed 18/10/11.

Kilgarriff, A., P. Rychly, P. Smrz & D. Tugwell. 2004. The Sketch Engine. In G. Williams & S. Vessier (eds), *Euralex 2004 Proceedings*. Lorient: Université de Bretagne-Sud, p. 105-116.

Pustejovsky, J. 1995. *The Generative Lexicon.* Cambridge MA: MIT Press.

Pustejovsky, J., P. Hanks & A. Rumshisky. 2004. Automated induction of sense in context. In *Proceedings of the 20th International Conference on Computational Linguistics.* http://acl.ldc.upenn.edu/C/C04/C04-1133.pdf, accessed 18/10/11.

Putnam, H. 1970. Is semantics possible? *Metaphilosophy* 1(3): 187-201.

Putnam, H. 1975a. The meaning of 'meaning'. *Minnesota Studies in the Philosophy of Science* 7: 131-193.

Putnam, H. 1975b. *Mind, Language, and Reality: Philosophical papers.* Cambridge: Cambridge University Press.

Rosch, E. 1973a. Natural categories. *Cognitive Psychology* 4(3): 328-350.

Rosch, E. 1973b. On the internal structure of perceptual and semantic categories. In T. Moore (ed.), *Cognitive Development and the Acquisition of Language*. New York: Academic Press.

Sinclair, J. 1966. Beginning the study of lexis. In C. Bazell, J. Catford, M. Halliday & R. Robins (eds), *In Memory of J. R. Firth*. London: Longman.

Sinclair, J (ed.). 1987. *Looking Up: An account of the COBUILD project in lexical computing*. London: Collins.

Sinclair, J. 1991. *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.

Sinclair, J. 2004. *Trust the Text: Language, corpus and discourse.* London / New York: Routledge.

Stubbs, M. 2001. *Words and Phrases: Corpus studies of lexical semantics*. Oxford: Blackwell.

Tesnière, L. 1959. *Éléments de Syntaxe Structurale*. Paris: Klincksieck.

Wray, A. 2002. *Formulaic Language and the Lexicon.* Cambridge: Cambridge University Press.