What Constitutes a Unit of Analysis in Language?[*]

Pernilla Danielsson (Birmingham)

Abstract

Over the last decade, the study of multi-word units has become increasingly popular and now these units seem to have reached a status where they cannot be ignored. This paper should be seen as a recapituation of the discussion around multi-word units, especially focusing on corpus evidence of such units and what is perceived to be relevant findings. An unsupervised method for extracting multi-word units from corpora is presented and the findings examined. However, rather than evaluating the results, this article will raise the question of what constitutes a 'good' multi-word unit? The article does not claim to give any conclusive answers but perhaps instead posing a few relevant questions.

1 Introduction

In this paper, I shall focus on a method for identifying multi-word units of analysis in language that satisfies two criteria: the method must be wholly automatic, not requiring human intervention except in evaluating the output; and the units so identified must be acceptable as genuine units of meaning (see also Danielsson 2001; 2003). The method will be illustrated using large general corpora of present-day English, the British National Corpus and the Bank of English.

It is by now well known that for the most part meaning belongs to multi-word units (mwus) rather than to individual words. For example, although the English word *scruff* is defined in the *Collins English Dictionary* as '*the nape of the neck*', its most frequent usage in the Bank of English is illustrated in the set of concordance lines below:

```
  is picked up by the scruff of the neck and
taken the game by the scruff of the neck with
 took the game by the scruff of the neck.
 drag Marti in by the scruff of the neck and
take the match by the scruff of the neck by
 Gavin Peacock by the scruff of the neck into
 take the game by the scruff of the neck, no
take the match by the scruff of the neck and
 grab the game by the scruff of the neck.
 took the game by the scruff of the neck. He
aken the match by the scruff of the neck in
```

Example 1. Concordance of the word 'scruff' from the Bank of English

Even without any further sorting, these concordance lines make it clear that the word *scruff* is mainly used in English in a set phrase, namely '*by the scruff of the neck.*' Furthermore, the verbs on the left hand side of the phrase tend to be forms of *take, grab, drag (in)* or even *pick up*. Together, these verbs seem to form a group denoting the action of grasping something (with your hands). *However, f*ocusing on the lines with these verbs, we find that the objects

[*] I would like to thank Susan Hunston, John Sinclair, David Oakey and Jordi Ginebra for valuable comments during the preparation of this document.

associated with *take by the scruff of the neck* are most often a *game,* a *match* or even an *opportunity*. In fact, in the Bank of English corpus there is only one occurrence in the corpus of a dog being picked up by the scruff of his neck, in which case *scruff* does refer to '*the nape of the neck*' as it is defined by CED. All the other occurrences have the different usage illustrated above, which only becomes obvious when viewing the repetitive patterns of language data through a concordance program.

There has been a huge growth of interest in phenomena such as this in recent years. Although it has long been known that words are distributed non-randomly in text, and that this non-random distribution carries information about meaning (Firth 1951/1957; de Saussure 1918/1959), it is only more recently that linguists have taken the implications of this seriously and have extended their investigations into multi-word units beyond the traditional categories of phrasal verbs, nominal compounds, and idioms. The popularity of the term 'pattern', for instance, is evidenced by the many corpus studies that include 'pattern' in their title (for example, Hunston/Francis 1999; Partington 1998; Hoey 1991). Interest in multi-word units has also grown, but with this popularity the number of ways to refer to them has also increased. References to *lexical items, units of meaning* (Sinclair 1996), *formulaic sequences* (Schmitt 2004; Wray 2002), *multi-word expressions* (ACL proceedings 2004), semi-preconstructed phrases (Sinclair 1991: 110) and *prefabricated units* (Erman/Warren 2000), can be used to show how heterogeneous the field is. The lack of uniform terminology may be just one of many clues to the confusion, and to the need for further research, in this area. Although the importance of multi-word units is accepted, there are no accepted answers to simple questions such as "What exactly constitutes a multi-word unit?" or "Where does a multi-word unit begin and end?"

The identification of mwus poses a problem for corpus linguistics and computational linguistics alike, in that whereas recurring sequences of words can be identified easily, such sequences are unlikely to coincide exactly with what a human researcher would accept as a 'unit of meaning' in a language. Unless this problem is addressed, computational linguistics in particular will find it difficult to embrace insights into language that prioritise meaning rather than structure (e.g. Sinclair 1991; 2004). As many of the applications of computational linguistics, such as text-mining, also prioritise meaning, this amounts to a missed opportunity.

There are in existence a number of algorithms for the identification of multi-word units, including for example Piao et al. (2005), Mason (2006), Smadja (1989, 1993) and Diaz et al. (1999, 2005). All, however, face difficulties because of the mismatch between what computer programs can easily do and what mwus are like. Firstly, recurring sequences of words can be identified if their length is specified in advance, as when bigrams or trigrams are extracted from a corpus. There is no reason to suppose, however, that these preset sequence lengths will map on to actual language units, and indeed as research continues it becomes apparent that there is no length specification to units of meaning, so that setting a sequence length in advance becomes counter-productive. Secondly, recurring sequences by themselves do not allow for discontinuous sequences, although it is well known that most units (e.g. *keep an eye on something*) allow for interpolation (e.g. *keep a watchful eye on something*) among other modifications (see Moon 1998). Finally, many computational approaches apply statistical significance measurements such as t-score, MI and log-likelihood, although it is uncertain what these statistics mean in the context of a non-random system such as language. Their application to the identification of mwus, therefore, is of questionable value.

The aim of the method described in this paper, then, is to identify units of meaning in monolingual corpora, assuming neither unit length nor sequentiality, and using only raw frequency as the statistical tool. The method will be outlined in the next section. The findings from this section will be used as a starting point for a brief discussion on whether human intuition and corpus data agree on what constitutes a multi-word unit.

2 Towards the identification of meaningful units

The methodology presented in this paper follows a very simple sequence of actions. First, all occurrences of the target node word (N) are identified in a large corpus and the most frequent collocate of the node (F1) is calculated within a span of 9 words (i.e. 4 words to the right and 4 to the left of N) 'Function' words are discarded. (They are identified by having an arbitrary cut-off point at a high position in the overall word frequency list for the corpus in question). Then all the lines which contain both N and F1 are selected, irrespective of the position of F1 relative to N. Taking just those lines, the most frequent word in them (apart from N and F1) is identified as F2. The process is repeated until occurrences of the new collocate identified fall to below 5 (an arbitrary cut-off point that may need to be revised). The procedure will now be illustrated using the word *jam* in the British National Corpus (BNC). First, all the concordance lines with the word *jam* as node are identified. From these lines the most frequent collocates of the node are calculated and the function words dismissed. The most frequent collocates thus identified are:

traffic
bread
jars
butter
strawberry
jar

Example 2. The list of the most frequent collocates around *jam*

For each of these collocates, concordance lines for the node and the collocate, where the collocate occurs within a span of 4+4 words are generated, starting with '*traffic*'.

```
  grimmest traffic jam seems better th
  man in a traffic jam who curses all
 nsport. a traffic jam in swansea.
n enormous traffic jam, and it looked
stuck in a traffic jam with your pulse
stuck in a traffic jam or being promo.
 ise every traffic jam will gradually
stuck in a traffic jam with their en
tting in a traffic jam with such a car
iting in a traffic jam. that driver
 alter the traffic jam in the slightest.
```

Example 3. Concordance for *jam* and *traffic*

Although the relative positions of *traffic* and *jam* were not specified, it is already clear that the most common position of *traffic* is one to the left of *jam.* A new calculation is now made, this time including all words, to find the most frequent collocate (F2) in the lines that include both *traffic* and *jam.* It is the word '*a*'. The concordance lines including the node word, *jam*, the first collocate, *traffic*, and the second collocate, *a* are now generated.

These concordance lines in turn are processed to identify the next most frequent collocate, excluding the words already selected. The new collocate (F3) is *in.* As there are still more than 62 concordance lines, we may continue to find yet another frequent collocate within these lines. The word *stuck* occurs 9 times. Here, we ran out of data to continue, as no other collocate occurs frequently enough to be included. Example 4 shows the concordance lines that include the node *jam* and F1, F2, F3 and F4: *traffic, a, in* and *stuck.*

```
stuck in a traffic jam the other day.
stuck in a traffic jam with your pulse raci
stuck in a traffic jam or being promoted.
stuck in a traffic jam with their engines
stuck in a traffic jam, and can't let
stuck in a traffic jam in the back of
stuck in a traffic jam, however , and
stuck in a traffic jam, you might reflect
stuck in a traffic jam? not a lot
```

Example 4. Concordance for 'stuck in a traffic jam'

From here, no other collocates occur with sufficient frequency to reach the cut-off point, and we may claim to have achieved the maximal unit based on the distribution in this corpus. Once the collocates are gathered, the next step is to order them sequentially. In this case it has been clear from the start that they will only present themselves in one order '*stuck in a traffic jam*', however, other units may allow more variation in sequence.

When a unit and the order of words has been established, the next step is to test for variations. Two types of variation may be identified: syntagmatic and paradigmatic variations. The syntagmatic variations come in the form of modifying words in between each of the words in the unit, as in '*a stroke of luck*' and '*a great stroke of luck*'. In the worked example here, intuition suggestions alternatives such as 'stuck in a hellish traffic jam'. However, these are not present in the BNC corpus so in the case here, we have no syntagmatic variations.

For paradigmatic variations, each word in the unit is tested to see if the unit allows for alternatives. For example, the corpus is searched to identify which other words can be found in the position of '*stuck*'when followed by the exact phrase '*in a traffic jam*'. The corpus offers *stuck, sitting, waiting,* and *caught.* In this particular context, the words seem to be related and offer a set of verbs that create a feeling associated with the annoying event of being held up in traffic. It should be noted that this association is context-dependent: in other contexts, *stuck, sitting, waiting* and *caught* would not be seen as a set of related words.

The next word to test for paradigmatic variation is *in*. Here, no variation is apparent. In the sequence '*stuck x a traffic jam*', *x* is always *in*. The same is true of '*a*'. Interestingly, it appears that this is most often the case for the high-frequency words. They are often said to provide structure rather than meaning, and as such they seem to object to modification and variation.

There are no variations for the word *traffic* either, , but the last variation test, for the word *jam* itself, gives a few alternatives, namely *nightmare* and *queue*. Again, these are two words which are not normally seen as forming a set with *jam*, but in this unit they count as viable variations.

The methodology presented here has previously been illustrated in Danielsson (2003) and was first introduced in Danielsson (2001). Here, it has been altered in some ways. In previous publications, the downward/upward distinction of collocates was used with the cut-off point set at the frequency of the node word. In this study, the cut-off point is set at a higher frequency. The original argument for using the frequency of the node word as the cut-off point lies in the assumption that less frequent words carry more meaning. Hence any collocate appearing less frequently than the node word can be expected to carry more meaning. Such an approach assumes that the entire vocabulary of a corpus will be run through the software; certainly a lengthy process. This has here been replaced by a simpler assumption, making an arbitrary cut-off at a point which may symbolise the distinction between function words and content words. For further discussion on this matter, see Danielsson (2001).

3 *Jam*: a few case studies

This section will be used to investigate further some of the units retrieved by the methodology introduced above. Due to space limitations only a few will be mentioned. We will begin by looking at units including the F1 collocate '*jar/s*', as this word occurred both in singular and plural form in the list in example 2.

From the OED (1999), we find that one definition of *jam* is:

> *A conserve of fruit prepared by boiling it with sugar to a pulp*
> *The action of jamming, the fact or condition of being jammed, or tightly packed or squeezed, so as to prevent movement; a crush, a squeeze*

This definition might be considered relevant in concordance lines that including *jam* and its F1 collocate *jar*.

```
   catching bees in a jam jar, fly swatting
 re was a candle in a jam-jar on the locker.
   from a candle in a jam jar to electric lic
  plonk a candle in a jam jar, it is worth ta
 of wild flowers in a jam jar. There are slate
    Wild Flowers in a Jam Jar). And 18 months
  and place them in a jam jar, porcelain bowl,
  nd puts them in her jam jar. She has about
  ching tiddlers in a jam jar at the canal.
  hornet trapped in a jam-jar. ROC's `Dead
 bit like a wasp in a jam jar. Davis: But
g up like a wasp in a jam jar." When I went,
petals and water in a jam jar. It'd be worn
 spit from a zoo in a jam jar. He said: 'This
```

Example 5. Concordances of 'jam jar'

Interestingly, however, in the *jam jar* concordance above, the meaning of *jam* as defined in the OED has disappeared. There is no reference to the actual conserve and this would make it difficult for a computational linguistic algorithm, based on word-sense disambiguation, to link the word *jam* to any sense of *jam*. *Jam jar* is mentioned in the OED but only as '*a container designed for holding jam*'. In the examples above, the container holds many things but never *jam*. *Flower*, *candles* and *small insects* are the more common contents of the jar.

The corpus data also throws up a few lines with *like a wasp/hornet in a jam jar*. These phrases refer to a specific sound and are in fact often preceded by the verb *sounds*. The physical object *jam jar* is now not a high priority, as as the phrase evokes a specific noise rather than a named insect inside a named container. It might be argued that it is the phrase *sounds like*, rather than *like a wasp in a jam jar* that is associated with a noise but a search into the corpus for the sequence *sounds like* tells a different story. This phrase, *sounds like,* is usually followed by *ideas* and *dreams*, *nightmares* or *a load of rubbish*; these are not normally things that give off a noise of any kind, as exemplified below.

Sounds like
a good idea
a lot of work
a load of [rubbish, rhubarb, gobbledegook]
a recipe for [disaster, boredom]
a [dream, nightmare]
a contradiction in terms

Example 6. The most frequent usages of the phrase 'sounds like' from the Bank of English

Returning to the target word jam, other identified multi-word units are: bread and jam, butter and jam, strawberry jam, spread jam, apricot jam, a jam sandwich, raspberry jam, toast and strawberry jam, toast and jam, jam and cream, jam or marmalade, jam tarts, blackcurrant jam, pot of jam, cake with jam, jam sandwiches, empty jam jars, tomato jam, and jam session.

At a first glance, the list may look a complete mix of various usages, but the list lends itself to further organisation. Many of the phrases follow the pattern of a word indicating a fruit or vegetable followed by jam

Strawberry jam
Apricot jam
Raspberry jam
Blackcurrant jam
Tomato jam

Further units like *jam session* refer to music performances which are often described as *impromptu* in the corpus, and which are again far from the fruit conserve meaning of *jam.* Others such as *bread and jam*, *butter and jam*, *toast and jam*, are however directly linked to the conserve meaning. Interesting observations about these binomials are that they tend to be stable in one order and only rarely are found in the opposite combination (*jam and bread* or *jam and butter*). This is a characteristic of many binomials (think of *knife and fork* for example, which you would hardly ever hear referred to as *fork and knife*) regardless of whether or not the order can be said to be logical.

4 Discussion

Although modern linguistics and computational linguistics both acknowledge the importance of multi-word units in language, these units are often still seen only to substitute words as units in specific cases, and therefore they are treated as instances of the same categories as words. For example '*in order to*' is treated as a preposition, and '*estate agent*' as a noun. These types of units would only suffice to cover some of the findings above. If mwus are restricted to those that resemble single words in this way, the full potential of the larger units such as *stuck in a traffic jam* is in danger of being ignored. This tends to happen if the word is still regarded as the 'real' unit of language analysis, with mwus as a special case, instead of accepting mwus as the norm, with the independent word, where necessary, being treated as the special case.

The question of evaluation remains: how should the results of this method of extracting multi-word units be assessed? Possible evaluators would be the linguist (are these relevant units of language?), or the lexicographer (would you put these in a dictionary?), or the developer of a computational language analysis tool (would you have these units in your lexicon?) The answers to all these question would probably be 'no'.

In traditional linguistics a word is viewed as having inherent meaning, and if discussing ambiguity in language the tendency is to only look upon the cases where ambiguity needs to be resolved, i.e. where the meaning of each word differs from the one that is said to be the word's inherent or prototypical meaning. However, there is a severe problem with this approach as, apart a few concrete nouns, words rarely have a clear prototypical meaning, evidenced by the differences of the primary sense definition of words in many dictionaries (especially when comparing between dictionaries which lists sense according to frequency, such as the COBUILD dictionary, and more traditional dictionaries of the English language).

It may be argued that in cases such as '*strawberry jam*', the multi-word unit has not altered the meaning of the prototypical *jam* (a conserve of fruit) in any way. Yet, what *strawberry* has done to this unit is to clarify, it has told us that out of all possible usages of *jam*, only one is relevant here. Most certainly this will be the case for very many of the words in corpus-extracted multi-word units. Their meaning has not altered, but it has become more specific. This specificity is most important in a larger system. When a particular meaning is expressed it will have an effect on the whole environment. Knowing that *jam*, when preceded by *strawberry*, refers to the conserve may affect the understanding of other units used at other points in a text. Hence, what may seem to be adding only little meaningful value in the term of a unit of meaning, may bring important information to the understanding of the full text.

But how do we comment on units such as '*bread and jam*' or '*toast and jam*'. If we are to fully embrace and understand the role of words in language, we may need to be aware of the possibility that the construction of multi-word units has much more to do with conventionalization than anything from the grammatical system. One interesting claim comes from Erman and Warren (2000), stating that at least 55% of the words in a modern-day English text are parts of some form of prefabricated multi-word units. As there is no, or only very limited, possible substitution of the words within the unit, there is little relevance in splitting them up into single words for further lexical or grammatical analyses.

The research by Erman and Warren (ibid) confirms what many language professionals have argued for years: in order to sound fluent a learner needs to know the conventional patterning around the words and not just the meaning and grammatical class of the word. Taking this into the world of computational linguistics, we may argue that any computational linguistic system that wants to recognize or synthesize language will need to have vast amount of information about the patterning. This may need to go further than the intuitions of language users. In a recent test of human annotators of multi-word units undertaken at the University of Birmingham showed that the units marked up by humans were shorter (usually only two words) than those retrieved by computer. Humans also tended to identify only multi-word units with clear lexical function, such as *run out of* and *last night*, whereas the computer would also retrieve longer units such as '*were reduced to ten men*' when given the same text (Danielsson forthcoming). This poses the question whether the mental concept of a multi-word unit has a restriction in length, and subsequently if such a length restriction is realized in the actual language Do we as humans have an intuitive understanding of what constitutes a unit of meaning, or do we need to be "trained" to see it?

This paper does not intend to answer all the questions associated with multi-word units but may serve to raise more questions. In particular, it questions the efficacy of using language users as the arbiters in evaluating the performance of mwu extraction methods. Perhaps for the time being we should instead acknowledge the corpus findings as proof of something that is important in language, based on the argument that if it were not important it would not be found repeatedly in a corpus. Such an attitude would alter the questions from '*is this unit a proper unit of meaning*?', to '*if this is an important unit in language, what effect does it have on our description of the overall language system?*'

References

Barnbrook, Geoff/Danielsson, Pernilla/Mahlberg, Michaela (eds.) (2005): *Meaningful Texts. The Extraction of Semantic Information from Monolingual and Multilingual Corpora*. London.

Danielsson, Pernilla (2003): "Automatic extraction of meaningful units from corpora". *International Journal of Corpus Linguistics* 8, 1.

Danielsson, Pernilla (2001): *The Automatic Identification of Meaningful Units in Language*. PhD Thesis. Göteborg University.

Dias, Gaël/Madeira, Sara/Pereira Lopes, José G. (2005): "Extracting concepts from dynamic legislative text collections". In: Barnbrook, Geoff et al. (eds.) (2005): *Meaningful Texts.*

The Extraction of Semantic Information from Monolingual and Multilingual Corpora. London: 5–16.

Dias, Gaël H./Guillore, Sylvie/Pereira Lopes, José G. (1999): "Language Independent Automatic Acquisition of Rigid Multiword Units from Unrestricted Text Corpora". *Traitement Automatique des Langues Naturelles (TANL)*: 333–339. Available at *http://www.di.ubi.pt/~ddg/publications/taln1999.pdf*, accessed July 18th, 2007.

Erman, Britt/Warren, Beatrice (2000): "The idiom principle and the open choice principle". *Text* 20, 1: 29–62.

Firth, John R. (1951/1957): "Modes of Meaning". *Papers in Linguistics 1934–1951*. Oxford: 190–215.

Hoey, Michael (1991): *Patterns of Lexis in Text*. Oxford.

Hunston, Susan/Francis, Gill (1999): *Pattern Grammar. A corpus-driven approach to the lexical grammar of English*. Amsterdam.

Mason, Oliver Jan (2006): *The automatic extraction of linguistic information from text corpora*. Ph D Thesis. University of Birmingham.

Moon, Rosamund (1998): *Fixed Idioms and expressions in English*. Oxford.

Partington, Alan (1998): *Patterns and Meanings*. Amsterdam.

Piao, Scott S. L. et al. (2005): "Comparing and combining a semantic tagger and a statistical tool for MWE extraction". *Computer Speech and Language. Special issue on Multiword expressions* 19, 4: 378–397.

Saussure, Ferdinand de (1918/1959): *Course in General Linguistics*. London.

Schmitt, Norbert (ed.) (2004): *Formulaic Sequences. Acquisition, Processing and Use*. Amsterdam.

Sinclair, John (1991): *Corpus, Concordance, Collocation*. Oxford.

Sinclair, John (1994): "The Search for Units of Meaning". *Textus* IX: 75–106.

Sinclair, John (2003): *Reading Concordances*. London.

Smadja, Frank (1993): "Retrieving Collocations from Text. Xtract". *Computational Linguistics* 15, 1: 143–177.

Smadja, Frank (1989): "Lexical Co-Occurrence. The Missing Link". *Literary & Linguistics Computing* 4, 3: 163–168.

Wray, Alison (2002): *Formulaic language and the lexicon*. Cambridge.

Dictionaries

Sinclair, John (1987/2001): *Collins COBUILD English Dictionary for Advanced Learners*. Third Edition. Glasgow.

Simpson, John A./Weiner, Edmund S. C. (1989): *The Oxford English Dictionary*. Second Edition, CD-ROM, Oxford/Ney York.

Butterfield, Jeremy et al. (2003): *Collins English Dictionary*. 6th revised edition. Glasgow.

Corpora

The Bank of English. A 450 million word corpus of English. Available at *http://www.collins.co.uk/books.aspx?group=140*, accessed July10th, 2007.

The British National Corpus (BNC). A 100 million word corpus of English. Available at *http://www.natcorp.ox.ac.uk/*, accessed July10th, 2007.