From lexis to syntax: the use of multi-word units in grammatical description

Oliver Mason [1]
The University of Birmingham

Abstract

We describe an approach to the description of sentence structures based on a linear model. The sentence is segmented using automatically identified multi-word units from a large corpus; recurrent elements from the corpus are matched up with fragments of the sentence. After positioning the current work in relation to recent related research we present two sample analyses and discuss the usefulness of this approach to syntactic description and possible applications.

Keywords : linear grammar, phraseology, lexis/syntax interface.

1. Introduction

Attempts to describe language predominantly assume a hierarchical model, such as a phrase structure tree. However, there are exceptions in Brazil's *Grammar of Speech* (1995), Hunston and Francis' *Pattern Grammar* (2000) and more recently in Sinclair and Mauranen's *Linear Unit Grammar* (2006). Other more lexically focused descriptions such as Hoey's *Lexical Priming* (2005) also discuss the syntagmatic relationships between words in a linear, non-hierarchical way.

Most approaches to syntax are based on assigning word classes to the tokens in a sentence and then applying a number of constraints (eg in the form of phrase structure rules or unification conditions) to the sequence of class labels to rule out certain arrangements. The principal weakness of such approaches is that they have to rely on the system of word classes to capture the regularities of lexical behaviour; but Sinclair (1991) has shown, for example, that even the most frequent member of the class 'preposition' (*of*) behaves quite unlike most other members. Similarly, many words occur only in a restricted environment, and cannot readily be replaced by other words of the same class. Sinclair thus introduces the 'idiom principle' which more adequately describes such usages, unlike the 'slot-and-filler' or open choice model, which works in some cases, but is generally not applicable.

As an alternative, more lexical information can be used for the description of sentence fragments. Examples are local grammars (eg Sinclair and Hunston 2000) and pattern grammar (eg Hunston and Francis 2000), where stretches of text are described through a combination of word classes, phrasal categories, and lexical items. Linear unit grammar (Sinclair and Mauranen 2006) completely abandons word classes and deliberately uses a pre-theoretical notion of 'chunk' for segmenting discourse into larger units, which are then assigned functional labels. Local grammars as described by Gross (1993) focus on (semi-)fixed expressions, but can be used to express larger patternings in sentence construction. As a formalism (based on finite-state machines) it can easily be applied to pattern grammar, and is also related to Brazil's basic method of analysing an utterance as a (finite-state) chain.

[1] Department of English, The University of Birmingham, O.Mason@bham.ac.uk

On the lexis-side of the traditional syntax/lexis divide is work on phraseology (eg Hunston 2001, Hoey 2005) which looks at the typical patterns in which words tend to occur. These patterns are specific to a particular word, and even closely related words (such as near-synonyms) rarely if ever occupy the same patterns. This contrasts sharply with the generalised slot-filler model of syntax based on word classes. Treating words as unique increases the precision of the description, but at the same time prevents any generalisations based on class-behaviour, hence the lack of any large-scale analyses.

In this paper we will try to bridge the gap between lexis and syntax from the lexical side, extending previous work on phraseology (Mason 2005) towards the description of sentence structure. The starting point is a set of multi-word units derived automatically from a corpus; these units are then matched with the sentence under investigation.

2. A Different View of Grammar

Traditionally, grammarians face several problems when describing the structure of language. The first, and most fundamental, is that they usually presuppose shared knowledge about a language, common to all its speakers, which clearly cannot exist in that form. Although we cannot yet investigate the content of various speakers' brains, we can surely exclude the possibility that somewhere an entity 'grammatical knowledge' exists which is identical for all members of a language community. Instead, each speaker of a language has their own grammar, derived from the individual's linguistic experience. These grammars will obviously be similar and to a large extent overlapping, but they will not be identical. Therefore any unified grammatical description has to remain an approximation based on the pool of shared linguistic knowledge of a speech community. This is a consequence of the view of language not as a single entity, but a conglomeration of a number of similar idiolects.

The second problem is that grammar traditionally uses the sentence as the fundamental unit of analysis. But sentences do not exist in spoken language, only in written texts, which are usually carefully edited following the artificial conventions of the writing system used. This leads to a restricted view of language, based on misleading notions of 'correctness' or grammatical well-formedness, which are of secondary importance given the purpose of language as a means of communication.

Formal approaches to the description of sentence structure furthermore take for granted a hierarchical (phrase) structure, going back to Bloomfield's notion of immediate constituents. However, language is not produced in that way, but instead is a linear sequence created in stops and starts. A hierarchical structure thus cannot account for the fact that the beginning of an utterance is already produced before the whole sentence has been completely worked out. Similar issues apply for the reception of language.

Unlike the hierarchical, a linear approach is more closely related to the way most language is received. Processing usually begins before a complete sentence has been heard or read, and quite often the remaining parts of a sentence can be predicted with high accuracy before its completion. In fact, in spoken discourse speakers often complete each other's sentences. This seems to make it unlikely for a hierarchal description to be appropriate for language, apart from carefully edited written sentences.

The grammatical description that we are trying to achieve is based on empirical principles, in other words we try to avoid any bias that could be introduced through pre-existing assumptions.

To summarise, we assume that

- each speaker has their own individual grammar derived from their linguistic experience
- written language is influenced by the norms and conventions of the writing system used
- a hierarchical system is not appropriate for describing the syntactic structure of utterances
- the structure of utterances is best described in a linear model

We will try to model the linguistic experience of a speaker through use of a general reference corpus, the British National Corpus (BNC). This will be our equivalent of Chomsky's idealised native speaker; however, as the BNC is clearly not a perfect match for a human being's linguistic experience we have to accept the limitations of this approach.

While principally our approach should work on spoken as well as written language, we will limit ourselves to written data initially. The main reason is that we require corpus data for the modelling of the user's language experience, and thus it is easier to fall back on the large corpora of written English available. That, however, does not mean that this approach should not work for either spoken language or languages other than English.

In the following section we will describe how we identify larger units from text corpora, before we proceed to analysing sentences with those units.

3. Multi-word Units

Sinclair and Mauranen (2006) decline any precise definition of 'chunk'; instead they state that dividing a text into chunks is something that comes naturally to fluent speakers of a language. While different speakers might choose different chunks, Sinclair and Mauranen have observed that there is usually sufficient overlap to postulate the existence of chunks as natural units of language. This is perfectly compatible with an individualist view of language, as each speaker will have their own individual language experience, and it would be illusory to assume perfect agreement on any such task.

However, it remains unsatisfactory that there is no objective way of deriving chunks, whose origin then remains in the realm of intuition. On the other hand, we do not want to impose any *a priori* constraints on the shape or form of chunks, such as grammatical well-formedness or the like. In the research described in this paper, we have thus chosen automatically identified multi-word units (MWUs) as preliminary chunks. These can be derived from corpus data in an objective way, and different corpora will lead to different MWUs, which is consistent with speaker variation. In the remainder of this section we will briefly describe the algorithm used for identifying MWUs.

In Mason (2006) we suggest two separate procedures to identify multi-word units, 'frames' and 'chains'. The latter are a variation on n-grams, where for a range of 'n' a fixed number of words is strung together, and then weighted using a length/frequency trade-off. The basic idea here is that short chains will be more frequent, but less interesting, whereas longer chains are more specific (and thus less frequent) and also more interesting. Frames, on the other hand, are based on the frequency of a word and its immediate neighbours: if the frequency of the words on either side of the starting word is higher than the word's own, they get added to the frame, otherwise the procedure stops. Here the underlying rationale is that lexical words carry information (as they are less frequent), whereas grammatical words are predominantly concerned with the arrangement of items in a sequence; an analogy would be a brick wall, where — when demolished — the mortar (grammatical words) will stick to the bricks (content words). Frames have been inspired by the collocational frameworks of Sinclair and Renouf (1991); only while Sinclair and Renouf look at a sequence of high-frequency items with a gap, eg *as — as*, we revert the perspective and start with the lower frequency content word and attach higher frequency words to it.

In the implementation used for the current research project, the sets of MWUs produced by the two procedures are merged, to provide a single source of MWUs for any given word. The list of MWUs is then filtered to exclude insignificant ones; in this case, an arbitrary threshold of 1% of the frequency of the most frequent MWU was chosen. As a result we have a list of MWUs for the word we started with. This is unlike other approaches (eg Stubbs and Barth 2003) which process a text at a time and extract MWUs that occur in the text but not those which are associated with a specific word.

The rationale for using the combined output of two different algorithms is as follows: *n*-grams are an established method for generating multi-word sequences, and they do not take into account any linguistic information that might bias the selection. By using sequences of different lengths and filtering them with a length/frequency trade-off, we are confident that the result

is both relevant (higher frequency) and interesting (longer sequence), while avoiding the extremes of very short high-frequency units or very rare long units. Frames, on the other hand, exploit basic information about a word type (its frequency of occurrence) and as such is more 'knowledge-based' than the straightforward *n*-grams. It introduces assumptions about the information structure of lexical items into the whole procedure, but through the combination with *n*-grams the effect of introducing a bias is mediated.

We will now apply the multi-word units retrieved from a corpus to the analysis of sentence structure.

4. Sample Analysis

We start off by looking at a sentence taken from the call for papers of this conference: *The papers presented at the conference will be available in proceedings on the first day.* For each word in this sentence we retrieve the multi-word units from the British National Corpus (BNC) as described in the previous section. We then select those units which match the surrounding words in our sentence and display the result in tabular form:

the	papers	presented	at	the	conference	will	be	available	in	proceedings	on	the	first	day
	PAPERS	presented	at											
	PAPERS	presented	at	the										
the	PAPERS	presented												
the	PAPERS	presented	at	the										
the	PAPERS	presented	at											
		PRESENTED	at	the										
	papers	PRESENTED	at											
	papers	PRESENTED	at	the										
the	papers	PRESENTED												
the	papers	PRESENTED	at	the										
the	papers	PRESENTED	at											
			AT	the	conference									
			at	THE	conference									
			at	the	CONFERENCE									
				the	CONFERENCE	will								
					CONFERENCE	will	be							
				the	CONFERENCE	will	be							
		presented	at	the	CONFERENCE									
						WILL	be	available						
						WILL	be	available	in					
						will	BE	available						
						will	BE	available	in					
							BE	available	in					
						will	be	AVAILABLE						
						will	be	AVAILABLE	in					
							be	AVAILABLE	in					
						will	be	available	IN					
							be	available	IN					
										PROCEEDINGS	on	the		
										PROCEEDINGS	on	the	first	
											ON	the	first	
											ON	the	first	day
											on	THE	first	
												THE	first	day
											on	THE	first	day
											on	the	FIRST	
												the	FIRST	day
											on	the	FIRST	day
												the	first	DAY
											on	the	first	DAY

Table 1. Sentence taken from Conference Call for Papers

The word in upper case is the respective 'node word', ie the word which was used as the starting point for the MWU extraction. Words in lower case are the associated context words which form a MWU with the node word. By tabulating the MWUs as we have done here it becomes apparent that they overlap and link up to form a longer sequence, similar to what Hunston and Francis (2000) describe as 'pattern flow'. We could say that the word *papers* prospects the following items *presented at the*, whereas *presented* reinforces the expectation of *at the*, *at* then prospects *the conference*, and so forth. 'Prospecting' is an important concept in linear grammar, as it restricts the choice of subsequent elements and thus leads us towards the idiom-principle, away from the open-choice-model.

Interestingly, *conference* then flows into *will be*, which could be classed as an instance of 'colligation' in the sense used by Hoey (2005): the word *conference* tends to occur frequently with expressions of futurity, in this case *will be*. Here we have a non-lexical notion (tense) which

could be realised in different ways.

There is only one point in this sentence where the flow of MWUs is interrupted, between *in* and *proceedings*. Here we can hypothesise the existence of a higher-level unit boundary, which is not crossed by the MWU chunks.

Even though we speak of *at* prospecting *the conference*, we need to be careful about the scope of such statements: they only apply to the analysis of an utterance, not its creation. While we could undoubtedly generate natural-sounding utterances by randomly stringing together overlapping MWUs, we would ignore the semantic aspect and the utterances would not be comparable to authentic ones. But in the analysis we presuppose that the utterance we are looking at is meaningful, so that the semantic dimension is implicit. There are certainly many more (in fact, 565 in total) MWUs that begin with *at*, but out of these that particular one has been chosen.

Looking back at the table, we can see that the MWUs we found in the BNC fully cover our sample sentence. This is in line with one of the principles mentioned by Stubbs (1993:2), "much language use is routine". Especially in fairly standard situations (such as giving information about conferences), we do not need to be creative. On the contrary, going back to routine usages we make it easier for the recipients to understand what we are saying, as it involves less effort to process something that one has already encountered before.

We could thus assume that the degree of MWU coverage changes according to the text type: texts which are easy to read ought to be described better using MWU chunks than highly creative ones or those which are more difficult to read. This obviously has to take into account other considerations, such as topic: since we model the speaker's linguistic experience through a general reference corpus (the BNC), texts which make use of specialised vocabulary will clearly have less coverage. But from a theoretical point of view this poses no problem, as the BNC is only an approximation in the first place. If we were to analyse an academic article, then we would get a higher coverage if we used a corpus of academic language for the retrieval of MWUs. This is consistent with the notion of a separate speech community, that of academics, which have separate shared linguistic experiences from other communities.

In a compressed format we can represent the outcome of the MWU matching procedure as follows: words which are not covered by an MWU are put in brackets, and places where MWUs do not overlap are indicated by a vertical bar. For the sentence above we would get the following:

> the papers presented at the conference will be available in | proceedings on the first day

All words are covered, and there is no overlap between *in* and *proceedings*. What we lose, however, is the information about which words are part of the same MWU, and which items are particular 'bridges' connecting different MWUs.

The next sentence we will investigate is from a children's book, *My Grandmother's Clock* by McCaughrean and Lambert. The analysis is given in two separate tables due to its length.

In this example we can see partial coverage only; that suggests that while fragments of the sentence are constructed according to the idiom principle, there are limits to it. One obvious point is the use of proper nouns, such as *King Zog*, which is clearly too specific to occur in a recurrent MWU, and would suggest that there could be a category 'NAME' which would allow for more flexibility in the recognition of MWUs; this could be accomplished by using Gross-style local grammars of for example names as a pre-processing step before the MWU recognition. Other candidates for open categories would be numbers and dates, however, we need to be careful that we do not introduce the slot-filler model through the back door.

We can clearly see from the above analysis how *opened* leads to *door*, and *door* to *in the front of the*. We continue with *to find out why*, which is another segment that incidentally coincides with a traditional unit, an infinitive complement. Then we have another break, at what would be called a clause boundary, where the following unit is introduced by a conjunction. Another gap is followed by a segment where *walking* and *stick* clearly are commonly used together, followed by a further (unspecified) element, in this case realised by *a picture of*. Here we can see that the

once	i	opened	the	door	in	the	front	of	the	clock	to	find	out	why
		OPENED	the	door										
	i	OPENED	the											
	i	OPENED	the	door										
		OPENED	the	door	in									
		OPENED	the	door	in	the								
		opened	THE	door										
		opened	the	DOOR										
	i	opened	the	DOOR										
				DOOR	in	the								
			the	DOOR	in									
			the	DOOR	in	the								
					IN	the	front							
				door	IN	the								
					IN	the	front	of						
			the	door	IN	the								
			the	door	IN									
					IN	the	front	of	the					
						THE	front	of	the					
						THE	front	of						
					in	THE	front							
					in	THE	front	of						
							FRONT	of	the					
						the	FRONT	of	the					
						the	FRONT	of						
					in	the	FRONT							
					in	the	FRONT	of	the					
					in	the	FRONT	of						
							front	OF	the					
						the	front	OF	the					
						the	front	OF						
							front	of	THE					
						the	front	of	THE					
					in	the	front	of	THE					
								of	the	CLOCK				
									the	CLOCK	to			
											TO	find	out	
											TO	find	out	why
											to	FIND	out	
											to	FIND	out	why
												FIND	out	why
											to	find	OUT	
											to	find	out	WHY
												find	out	WHY

Table 2. My Grandmother's Clock, *first part*

MWU *stick and a* stops short of specifying the following noun, and the subsequent MWUs have no reference back to the previous one, as the combination of walking sticks with pictures is not a usual one. We have already commented on the final gap, where the subject of the picture is omitted. In the compressed notation the sentence can be represented as:

> (once) | i opened the door in the front of the clock to find out why | and there was nothing inside | (but one umbrella) | a walking stick and a picture of | (king zog)

While less complete than the previous analysis, we can easily find explanations for the gaps in the MWU flow. It is encouraging that some breaks coincide with units of traditional grammatical description, which indicates that despite general issues traditional grammar does indeed seem to reflect linguistic structures, even if derived in a different way.

However, it is difficult to map the MWUs directly onto grammatical units, or describe their exact function. In the kind of description we would achieve here we would have to abandon traditional syntactic structures, and approach a sentence from the word-level. We would be able to find linkages between individual words or groups of words, which has been exemplified in the above analyses. Obviously two sentences is only a very small data set, but further data analyses so far have produced similar results.

5. Discussion

What conclusions can we draw from these two sample analyses? One methodological issue is that we are looking at the utterance after it has been completed, but at the same time we talk about a word 'prospecting' another, which would only happen at the time of production. When we earlier, in the first sample analysis, said that *conference* occurs with expressions of futurity, then this obviously applies to this particular sentence. It does not preclude examples such as

and	there	was	nothing	inside	but	one	umbrella	a	walking	stick	and	a	picture	of	king	zog
AND	there	was														
AND	there	was	nothing													
and	THERE	was														
	THERE	was	nothing													
and	THERE	was	nothing													
and	there	WAS														
	there	WAS	nothing													
and	there	WAS	nothing													
	there	was	NOTHING													
and	there	was	NOTHING													
	there	was	nothing	INSIDE												
								a	WALKING	stick						
									WALKING	stick	and					
								a	WALKING	stick	and					
								a	walking	STICK						
										STICK	and	a				
									walking	STICK	and					
								a	walking	STICK	and					
											AND	a	picture	of		
												a	PICTURE	of		
											and	a	PICTURE	of		
											and	a	PICTURE			
												a	picture	OF		

Table 3. My Grandmother's Clock, *continued*

Plenary papers from the conference were published in Comparative Criticism 24 (2002), which shows a full flow from *papers* to *in*.

Language production involves making choices; what we can observe is the result of these choices. And work in the areas of collocation and colligation indicates that words and their syntactic contexts are co-selected, in other words that syntax and lexis cannot be separated. One further property that we can determine through the kind of analysis presented here is the proportion of routine *vs* creative use of language: the larger the coverage of the sentence, the higher the degree of re-use, and consequently a low coverage indicates more creativity. Arguably it is easier to process a sentence if it follows expectations, so one possible application for this procedure could be measuring the readability of a text.

Looking at the lexis-grammar interface, we can postulate that MWUs are suitable candidates for chunks in a grammar such as LUG (Sinclair and Mauranen 2006). Sinclair and Mauranen use the speakers' intuitive choices for the segmentation of text into chunks. The MWUs described here have one property which would make them unsuitable for such an application, namely that they are often overlapping. However, we could hypothesise that those cases where there is a gap in the MWU sequences reflects a definite chunk boundary, while overlaps show boundaries which are not unanimous. This is something that would have to be investigated further.

Brazil (1995) describes utterances in terms of chains of (grammatical) elements, whereas Sinclair and Mauranen (2006) look at them from the point of view of chunks to which they assign functional categories such as message-oriented or organisation-oriented. The modeling of structure as a sequence of (possibly overlapping) MWUs is not far enough developed yet to add an interpretation or functional description to the elements. As such it is still at the stage where elements are identified and the lexical relations between the elements which link them up can be investigated. In a way we are still closer to Hoey's idea of lexical priming as an influence on grammar than we are to a grammatical description along the lines of Brazil or Sinclair and Mauranen.

6. Conclusion

In this article we described a new approach to the description of utterance structure. Abandoning the idea of a hierarchical structure of sentences we instead opt for a linear model, similar to one developed by Brazil (1995), Hunston and Francis (2000), or Sinclair and Mauranen (2006). We have not yet introduced functional labels for the elements we identified, and due to the fact that those elements are identified in a fully automated process this will be less straightforward than if we had chosen those elements using our intuition instead. This is one area where this work will need to be expanded further. On the other hand, they might not be necessary if our grammar is predominantly based on the choice of lexical items and their corresponding MWUs.

One current application of our model is to specify the degree of creativity involved in the production of an utterance. By identifying fragments that have been used before in a large corpus we can account for recurrent usages, ie prefabricated units that are part of the 'standard' vocabulary of sub-phrasal units. Applying those units to a grammatical description would then be similar to data-oriented parsing (Bod 1998).

While this project is still at an early stage, it nevertheless comprises a promising approach to an empirical description of language. It also bridges the gap between syntax and lexis, by avoiding previous assumptions about phrase structure, and starting from recurrent word combinations which have been assembled following general principles rather than preconceptions about grammaticality.

References

BOD R. (1998), *Beyond Grammar: An experience-based theory of language*, Cambridge University Press (CSLI Publications).

BRAZIL D. (1995), *A Grammar of Speech*, Oxford University Press.

GROSS M. (1993), "Local Grammars and their Representation by Finite Automata", in Hoey M. (Ed), *Data, Description, Discourse*, HarperCollins, London : 26–38.

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

HUNSTON S. (2001), "Colligation, lexis, pattern, and text", in Scott M. & Thompson G. (Eds), *Patterns of Text*, John Benjamins, Amsterdam : 13–33.

HUNSTON S. and FRANCIS G. (2000), *Pattern Grammar: A corpus-driven approach to the lexical grammar of English*, Benjamins, Amsterdam.

MASON O. (2005), "Automatic Identification of English Multi-Word Units", in *Proceedings from the Corpus Linguistics Conference Series*, vol. 1.

MASON O. (2006), *The automatic extraction of linguistic information from text corpora*, PhD thesis, Department of English, The University of Birmingham.

RENOUF A. and SINCLAIR J. M. (1991), "Collocational Frameworks in English", in Aijmer K. & Altenberg B. (Eds), *English Corpus Linguistics*, Longman, London : 128–144.

SINCLAIR J. M. (1991), *Corpus, Concordance, Collocation*, Oxford University Press, Oxford.

SINCLAIR J. M. and HUNSTON S. (2000), "A Local Grammar of Evaluation", in Hunston S. & Thompson G. (Eds), *Evaluation in Text: Authorial Stance and the Construction of Discourse*, Oxford University Press, Oxford : 74–101.

SINCLAIR J. M. and MAURANEN A. (2006), *Linear Unit Grammar*, Benjamins, Amsterdam.

STUBBS M. (1993), "British Traditions in Text Analysis–From Firth to Sinclair", in Baker M., Francis G. & Tognini-Bonelli E. (Eds), *Text and Technology. In Honour of John Sinclair*, Benjamins, Amsterdam : 1–33.

STUBBS M. and BARTH I. (2003), "Using recurrent phrases as text-type discriminators", in *Functions of Language*, n° 1, vol. 10.