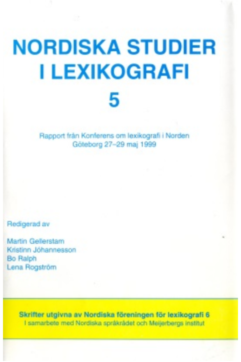

NORDISKA STUDIER
I LEXIKOGRAFI
5

Rapport från Konferens om lexikografi i Norden
Göteborg 27–29 maj 1999

Redigerad av

Martin Gellerstam
Kristinn Jóhannesson
Bo Ralph
Lena Rogström

Skrifter utgivna av Nordiska föreningen för lexikografi 6
I samarbete med Nordiska språkrådet och Meijerbergs institut

John Sinclair

Lexical Grammar

Introduction

This paper concerns the relation between the two types of pattern that are mainly recognised as the means whereby language creates meaning. The terms *grammar* and *lexis* will mainly be used for these, but instead of grammar you will sometimes find *syntax* or *structure*, and instead of lexis you may find *semantics* or *vocabulary*. But there is always this basic distinction, of a component which produces patterns of organization and a component which produces items that fill places in the patterns; the items tend to be chosen individually, and with little reference to the surrounding text.

The title of the paper is *Lexical Grammar* and not *lexico*-grammar. Lexicogrammar is now very fashionable, but it does not integrate the two types of pattern as its name might suggest – it is fundamentally grammar with a certain amount of attention to lexical patterns within the grammatical frameworks; it is not in any sense an attempt to build together a grammar and lexis on an equal basis.

When a dichotomy is firmly established in a culture, it is difficult to find a name for it or to talk about it as a unified whole and not two different things; that is the problem here. Recent research into the features of language corpora give us reason to believe that the fundamental distinction between grammar, on the one hand, and lexis, on the other hand, is not as fundamental as it is usually held to be and since it is a distinction that is made at the outset of the formal study of language, then it colours and distorts the whole enterprise. It is worth considering how far, using modern techniques, we can get in describing a language without resorting to such a distinction.

Grammar and Lexis

The distinction between grammar and lexis is a very basic model of language; there would be no motivation to reconsider it unless new evidence gave rise to concern about its accuraccy. One reason for such a model becoming so well established could be simply that before the computer age linguists were unable to describe all the complexity of language at once; since it could be represented as a framework and a set of choices to fit the frames, one of those elements could be held steady and the other varied against it. So we could forget, temporarily, about the patterns of semantic choice while we look at the organization of the structures; and then the process could be reversed, and when we come to look at the words and their meanings, then we do not consider at this point whether they are subjects or objects of clauses or objects of prepositions, if they are noun phrases, because that part of the overall organisation is suspended.

In other words, we can put forward for consideration the suggestion that this initial division of language patterning may not be fundamental to the nature of language, but more a consequence of the inadequacy of the means of studying language in the pre-computer age. When the linguist had nothing but his or her five senses, memory and internal awareness, it was difficult to analyse such a complex matter as language; consider phonetics, for example, before the sound wave could be slowed down and divided into its components. Without the ability to manipulate language externally, the observer/analyst has to leave some things steady, or hope they stay steady while other aspects of the whole are examined. And the problem is, in language, that they don't stay steady. So we should at least question the wisdom of dividing the meaningful patterns of language into two at the outset.

Abstractions

There is a related point to be made separately, but also a consequence of the position of the human observer. It is generally recognised that the meaningful patterns of language are of an abstract nature, which is one reason why they are so difficult to explain, and to use in teaching; from the perspective of grammar they are more abstract than they seem to be at first sight. It is possible that the reason for their unexpected level of abstraction is that grammar typically is realised through the common words and morphemes – that is, they seem to be familiar, but in fact many of them are multiply ambiguous and in a complex relationship with the categories that they realise (Sinclair 1999a). So grammar is superficially easy to observe but much more abstract than appears at first sight.

In contrast, the lexical patterns are very difficult to observe, because they are realised by a large vocabulary of infrequent words, and so it is not easy to work out the recurrent patterns that lie beneath the massive variation. The patterns are patterns of combination, and this compounds the problem; whereas in grammar the recurrence of frequent words makes it fairly easy to notice patterns of combination, in lexis the combinations had only been seen in a few hundred idiomatic expressions which were so remarkable that they had to be accounted for separately. With large corpora and powerful computers we are at the frontiers of a new view of language, where we can appreciate its full complexity without getting hampered by the detail.

It is thus no accident that linguists up till now have developed grammars much more than dictionaries and lexicons; we tend to have very elaborate grammars, which contain intricate apparatus with ranks and hierarchies and structures and all sorts of categories, with many different kinds of organization and in contrast we have very, very simple models of lexical structure, which are mainly one-dimensional, based on the word. There is an ad hoc set of terms for multi-word units like idiom and cliché and saying and proverb, but all these are ill-defined terms, and there is no other network of interconnections between one word and another.

Again this disparity in our descriptions does not necessarily reflect the nature of language, but rather it reflects our collective inability to process language with sufficient power and understanding to see that the complexity of the language as seen from a lexical point of view is just as great as the complexity of the language as seen from the grammatical point of view. So we may expect that simple artefacts like dictionaries will give way to more complex lexical architectures – indeed the development of dictionaries with an influence from corpus research has begun to move in this direction.

Meaning and Structure

There is one consequence of the initial separation of language patterning into two contrasted types that could be very important. To bring it out clearly we will use the terms *meaning* and *structure*. In brief, the point is that if we ignore the meaning while we are describing the structure, then of course we have removed the meaning and will not be able to get it back while we are focused on the structure. That is one way of expressing the problem of grammar, and it has been obscured from careful examination by a kind of meaning substitute. This is the curious terminology that we use, things like positive/negative, singular/plural, active/passive, and so on. If we look at them carefully, these terms are of course quite substantially inaccurate. ”Singular” does not always mean ”one,” and ”plural” does not always mean ”more than one”. ”Present” does not always refer to the time of the utterance, and ”past” certainly does not always mean some previous time. We have learned as part of our culture to suspend disbelief when we encounter these terms, and apply a rough criterion of *mutatis mutandis* to their interpretation; ”singular” means ”not more than one, if whatever it is is countable, otherwise general reference”. The point is that because these terms are not sensitive to the meaning, then they cannot actually be used directly to elucidate the meaning of text. The distinctions could have any labels at all, and at best they have a mnemonic function (this argument is well-supported by the retention in current grammars of terms like "finite" or "voice", which bears little relation to meaning in current English).

The only meaning that grammar provides is differentiation. From the *valeur* of Saussure to the systems of systemic linguistics and the choices of transformational grammars, then, the only way in which grammar creates meaning is by setting up mutually exclusive choices, and it exists purely as a record of the choice itself; the significance of the choice – whether a past tense verb relates to past time or present or future time or modality – is determined elsewhere.

If we now view the structure/meaning divide from the other perspective, and look at semantics without structure, then the typical way of presenting the meaning is the dictionary. A dictionary simply lists in an arbitrary order, which we call alphabetical[1] the items that it regards as being meaningful, which are usually the words of the language, and it tries to assign one or more meanings to each of the words. That is the characteristic model of a dictionary. The meanings are denied access to the structural organization that can put them together and show how they work. For example one meaning given in a recently-published dictionary for the word "white" is "counterrevolutionary, very conservative, or royalist"; if this meaning is still current it would take some ingenuity to specify the structural circumstances under which it could occur.

So therefore substitutes, again, are offered, this time standing in for the the linguistic organization that has been discarded. There are in semantics two major types of organization that have been imported; one of these is referential semantics, and the other is logical semantics; let us consider them in turn. The assumption of referential semantics is that meanings are organized with

[1] Alphabetical order is an order whose only virtue is that it is taught to all literate members of societies which use it. The fact that it is the only means of organising the vocabulary of a language merely emphasises the failure of linguists to find a better one.

reference to the world outside; words have meanings which can be understood by indicating objects, events and attributes in the world to which they refer; for abstract entities there is the "figurative" mode which works analagously. This is simple and seems to be broadly usable for a very large range of usable phenomena, and is widely used in education, but from a theoretical point of view it is absurd. Consider the proposal for a moment – on the one hand there is language, which we know is a highly organized phenomenon that operates under major constraints such as linearity, and on the other hand there is the world, which after thousands of years of research we still see as pretty chaotic, exceptionally complex and totally unable to be encompassed in a simple description. We are asked to accept that reference to the world can elucidate the structure of language? We have some reason to believe that language can elucidate some aspects of the world, but hardly the other way round. At best the referential links can help in for example supporting the acquisition of language by a child, before the child can cope with semantic abstractions.

The other type of imported semantic structure that is popular is logical semantics. It seems to have some advantages, being rigorous and much of it being quite close to the patterns of natural language (as well it might be, being derived from them). But it is crucial to the understanding of natural language that the organisation is *not* exact, and is not reliable as an indication of logical relationships. As with the definition of terms in grammar, there is again the problem of the partial fit, the inexact fit.

Here is a brief example to show the problems of relying on logical analogies too closely. Many commentators have noted that the "conditional" *if* does not always have its logical force, for example in the following instances culled from a large number of candidates in just one category of The BNC Sampler (spoken business):

> ...which is obtainable from Christian Aid if people want....
> I'm just thinking for the meeting if we could photocopy some Yes
> I'll be actually chairing the meeting for him. So so if you'd like to
> kick off...

> Mm. Yes. Mm. Erm Mm I could if I could just pick up one other point about you know...

And another one noticed casually in reading:

> If you believe me, I swung along that road whistling. (*The Thirty-Nine Steps*, John Buchan)

And one which has already occurred in this paper (one of two):

> If we look at them carefully, these terms are of course quite substantially inaccurate.

The Axes of Language Patterning

We now move our perspective to a closely related dichotomy that has long been recognised in language description – the two fundamental axes of language patterning, the paradigmatic and the syntagmatic. They are usually depicted as horizontal and vertical, with the syntagmatic axis on the horizontal, because the languages of modern Europe are written in horizontal lines, and the paradigmatic on the vertical.[1] The paradigmatic axis specifies the possible choices at a particular position on the syntagmatic axis, and the syntagmatic axis controls the structure which is being elaborated. So what we observe in language text is the syntagmatic; the paradigms are the total of what might have been chosen instead.

[1] The most available instance of this mechanism in some societies is the fruit machine, which used to be found in almost any public house in UK. A fruit machine consists of three revolving cylinders, each of which bears a number of drawings of fruit. The player pulls a handle at the side which causes the cylinders to spin independently of each other, and come to rest in a chance combination. A row of three fruit thus appears in a central grille, and if the row corresponds to one of those in a list on the side of the machine, (eg banana banana banana) then the player wins, and receives several times his or her stake. The central row corresponds to the syntagmatic axis, and the cylinders contain the paradigmatic choices; a well-formed structure is one of those that wins a prize.

Now, one of the interesting things about these two axes is that they cannot be simultaneously observed; you must hold one of them steady in order to look at the other. We shall return to this point, but no doubt this is the reason why we have had the division into grammar and lexis from an early stage. It is important to notice that the theoretical development of grammars in recent years has moved across this divide. If we were to map the "grammar" composite (including syntax, structure) and the "lexis" one onto the two axes, then the obvious pairing would be grammar on the horizontal axis and lexis on the vertical – a model of language often called the "slot-and-filler" model, the one presented at the outset of this paper. The syntactic structures form a series of slots, and these are filled with choices from the dictionary. The well-known models of transformational grammar are partly structured in this way, for example at the interface between the phrase structure and the lexicon, where the phrase structure specifes the features that any word must have in order to make a well-formed sentence, and the lexicon associates each word with a bundle of features. However, other influential models insist that they are primarily, if not exclusively, paradigmatic – notably Systemic-Functional Grammar (see Halliday 1995:15).

The syntagmatic patterns of language are not given meaning in a paradigm grammar, nor, of course, are they given meaning in a dictionary type of lexis. The syntagmatic patterns in a grammar are either offered as related through a common node, or they are simply declared. The syntagmatic patterns of lexis only appear in the byway of idiomatic phrase, where they are offered as joint realisations of a single meaningful unit, indicating that they have no meaning in themselves.

Let us consider the grammatical positions a little more. In phrase-structure rules like

$$S \rightarrow NP\ VP$$

the only relationship between NP and VP is that they are both derived from S in the same operation; their sequence is also determined in this single step. In the early days of generative grammar there was a plus sign in between NP and VP,

$$S \rightarrow NP+VP$$

but this signalled a quite spurious relationship pertaining on the syntagmatic axis, and became unfashionable.

Where syntagmatic patterns come into being by declaration, there is no explanation of where they come from or how they are to be deployed. The structure of an English clause is said to involve Subject-Predicator-Object-Adjunct, for example, but these categories are mutually defining, and do not have meaning until they are mapped into sets of choices, for example that a transitive clause is one without an object. So, neither in the study of the lexis of the language nor in the study of the grammar of the language are the syntagmatic patterns given meaning. This is to a great extent because there is no framework within which they can be shown to have meaning, because meaning is largely held to reside either in the grammatical choice – on the paradigmatic axis – or in the lexical choice of a word to deliver a meaning.

Syntagmatic Meaning

There is no effort, let us say in summary, to discover or create meaning on the syntagmatic axis; it is the responsibility of a paradigm grammar to build in all possible syntagmatic meaning as constraints on the paradigmatic choices. But such a venture would be remarkably complex, so in practice those grammars fail to describe carefully enough the *combinations* of choices that are just as central and meaningful and rule-governed as the single paradigmatic choices. They give tacit approval to the well-formedness of millions of sentences which range from the odd to the bizarre, and – by claiming as a series of choices phrases which we know to be a single choice – they claim large amounts of meaning which we know those choices do not create.

In corpus linguistics, by contrast, we have to work on the assumption that meaning is created on both axes; for want of more accurate information we may assume that they contain equal meaning potential. There is no reason why one should have a priority in meaning potential over the other. We assume a rough balance between what I have called (Sinclair 1996) the *phraseological tendency*, the tendency of a speaker/writer to choose several words at a time, and the *terminological tendency*, the tendency of language users to protect the meaning of a word or phrase so that every time it is used it guarantees delivery of a known meaning. As we get to know more, these assumptions may well be revised.

Above we have presented a model of language as a balance between opposing forces related to the two axes of language patterning, and above that is an assertion that the two axes cannot be simultaneously observed; these sound like good reasons for keeping them apart, and describing them separately. However, the argument of this paper is that if pattern and meaning are to be aligned, then the two axes have to be inter-related for as long as possible in the description. Consider, for example, the classic model where a choice is made on the paradigmatic axis, which will lead to a particular word appearing in the text. Now it is an axiom of the present approach to corpus linguistics that meaning and cotext are inter-related in such a way that involves at least partial coselection; so the knock-on effect of a paradigmatic choice will be felt on the syntagmatic axis. If we start from the other axis, then any existing or proposed pattern of choice on the syntagmatic axis provides a framework for the interpretation of any choice to be made on the paradigmatic axis.

Practical Consequences

The remainder of this paper gives some indications of the direction in which this argument is heading and the kind of consequences it is likely to lead to. First we will re-examine the nature of choice and meaning, then look further into the "meaningful" terminology of grammar, and finally pose a question about an important type of meaning that is largely ignored by both the grammatical and lexical traditions.

Meanings from Nowhere

Let us begin by revisiting the information-theoretic model of paradigm grammar, which says that choice equals meaning, that the number of choices determines the amount of meaning available in each case, and the precise positioning of the choice in the structural framework determines much of the type of meaning that will be created by the choice. A description within this model must take great care that each set of choices is actually relevant and applicable at each point. Because if it is not – if another factor in the environment is affecting the range of choices on offer, then unless the grammar is revised it is creating more meaning than is in fact available.

If this manufacture of illusory meaning is institutionalised throughout a complex grammar, there are two obvious consequences. One is that the grammar (and the grammarians) are misled into thinking that their apparatus is more powerful than it actually is; the other is that there is little meaning left over to be assigned by the lexical structure of the language. Now that it can be demonstrated by corpus evidence that a large proportion of the word occurrence is the result of co-selection – that is to say, more than one word is selected in a single choice – every time that this can be demonstrated there is one less item of meaning to be allocated to the grammar. If you have two words that are selected in the same choice, then they cannot be independently selected. Early estimates were that up to 80% of the occurrence of words could be through co-selections, which would leave, of course, only 20% for the sort of independent paradigmatic choices of the grammar. A recent paper by May Fan (1999) gave hard evidence for this in regard to one of the common verbs in English.

Let us work through a characteristic example. There is a phrase in English, a common recurrent phrase, ”out of the corner of my eye,” as in ”I saw something out of the corner of my eye.” There are seven words in the phrase, and they all simultaneously choose one unit of meaning, to do with peripheral vision. Within this primary meaning, there are one or two variants of individual words, and this is where the corpus is essential, because the intuition cannot be relied on: ”out of” can sometimes be replaced by ”from”, and ”my” is a possessive adjective that can have other, but probably only singular forms; people do not collectively see things out of the corners of their eyes, so I think ”their eyes” is going to be very unusual. This is the full extent of the variation associated with this phrase; the remaining words are fixed, and do not realise any choice beyond the first, overall choice of meaning. So neither of the occurrences of *of* above are the normal occurrence of the preposition, because *of* is fixed in this phrase,[1] and so are ”the”, ”corner” and ”eye”. The word ”my” can be alternated with other possessive adjectives. So here we have a seven-word phrase which realises one overall choice and at most two subsidiary choices. The choice between ”out of” and ”from” here is a stylistic choice rather than a choice that delivers a totally different type of meaning – there are not two different places, and ”out of” and ”from” are just different ways of expressing the same basic position.

These single choices can consist of seven words with ease; the phraseology of English quite frequently produces co-selections of five, six and seven words, and there are even some of up to twelve. In this connection Miller (1956) comes to mind. Miller showed that for most people the short-term memory handles seven items with ease.

[1] It would be a digression to argue here that *of* is not a preposition in such structures; for that see Sinclair 1991.

Cross-border categories

Corpus evidence consistently shows that the ways in which a meaning can be realised extend well beyond the definitions of grammatical categories. In pointing out above that grammatical terminology did not correspond to semantically coherent categories, we did not tell the whole story. Consider a term like "negative", which will contrast with "positive" in a two-term system of "polarity". There are a number of realisations of grammatical negatives in English, "no" and "not" and so on. There are also semi-negatives like "hardly" and "scarcely", which share a number of features with true negatives, but not all; these are not normally considered as grammatical negatives.

But there are also morphological negatives like the prefixes "un-" and "in-", which are not recognised in a clause grammar, so that "I am unhappy" is positive and "I am not happy" is negative. We also find that negation as a concept can be lexicalised, so that the verb "refuse" for example has a negative force; "he refused to go" is the same as "he would not go" and yet it is a positive clause in the grammatical sense.

It is easy to understand the grammarian's wish to keep negation pure and simple; to accept lexicalised negation is a slippery slope, and no-one knows what lies at the bottom of it. But if we are intent on elucidating the meaning of running text by analysis, then all these different ways of indicating negation are perfectly acceptable realisations, and if supported by corpus evidence we can take them all together, straddling the borders between grammar and semantics. This straddling is an important feature of lexical structure; lexis is not the residue of a grammatical description, but a different way of describing the same events; it is not bound by the conventions of grammar, and it can recognise a wide variety of realisations of meaningful choices.

The grammarian is left in a dilemma; the more sensitive grammars recognise that categories of meaning like "negative", "modal", "possessive" can readily be lexicalised – or to be more neutral, can occur in grammatical or lexical or morphological realisations – so a complex realisation route is devised for them.

The particular way in which they are realised is then of secondary importance compared with the primary creation of meaning, which is the operative process. The question must arise of the relevance of, for example, the grammatical choice between positive and negative to the study of meaning when negative meaning can be created in so many alternative ways; and, more fundamentally, how valuable is it to be able to point out that there are many clauses which are grammatically negative but in relation to meaning, positive, and vice versa?

Semantic Prosody

Another important point to be made in the study of lexical grammar is the emergence of many latent categories of meaning, which have not been recognised in published grammars, and only occasionally in the very latest dictionaries. The first to be noticed were of the type ”something nasty” or ”something worrying” or ”disturbing”; later others like ”something magnificent”, ”socially appropriate”, ”positively constructive” etc. These are showing up as repetitive categories that are neither completely grammatical nor completely lexical but are nevertheless very important from a structural point of view. So once again we have to allow for the meaningful categories not to be confined within the grammar as it is normally presented, and if we divide language into these two major categories, then we will never be able to get them satisfactorily together again; also we have to add that the grammar cannot be trusted to set up such essential categories of meaning because it is not sensitive to them.[1]

Here it has to be said that the perceptions of native speakers are not to be trusted either; the referential element in meaning is frequently assigned a priority over the attitudinal, for reasons that are not justifiable; clearly an awareness of both aspects of

[1] Here the new ”Pattern Grammars” (Hunston and Francis 1999) take the innovative step, guided by corpus evidence, of associating some of these meanings with structural patterning.

meaning is necessary for accurate deployment of the lexical item, and if this is not available it is arguable that more difficulty may arise from a mistake on the pragmatic side than on the referential. To give a real-life example, in the preparation of a dictionary for native speakers of English by the *Cobuild* team, there was a strong feeling among editors and publisher that whereas for learners of English it may be necessary to state the attitudinal meaning, this is already available to native speakers. So the *Cobuild* definition (1987) for *scrawny* is "unpleasantly thin and bony", while in *Today's English Dictionary* (1995) it is "thin and bony"; the two dictionaries define *prattle* identically but *Cobuild* adds "an informal word, often used showing disapproval".[1]

Word Class

Professional linguists should not be surprised to experience a rather disturbing effect from the massive surge in the availability of evidence and the growing sophistication of the tools for examining it and testing hypotheses against it that corpus linguistics has brought. Some of the vague but useful categories of traditional language analysis, which have served humans well for centuries, are not easily replicated in computational routines; for example "parts of speech" or "word class" labelling. Human beings have little difficulty assigning words to a dozen or so word classes, but machines have exposed just how untidy a categorisation this is. For English, which has had a lot of attention over many years, there is little or no consensus about how many labels there are – the variation from one analysis to another is very large – or how they are defined. The persistence of researchers has resulted in a significant movement of focus, so that the process is now called "morphosyntactic tagging" – in other words it was found necessary to use some syntactic

[1] This was one of the few arguments that I, as Editor-in-Chief of Cobuild, lost; but I am still puzzled at the conviction that native speakers may need to know the referential meaning of a word but not its attitudinal/pragmatic one.

information in order to complete what was originally a morphological analysis.

This movement of focus is well recognised in corpus linguistics – the need to examine the context of an item in order to determine its function or meaning. But nothing seems able to shake belief in the underlying assumption that all the words of a language naturally fall into a small number of classes. The information from a computer examination of a corpus suggests quite otherwise, as I have argued on several occasions.[1] Since few inflections survive into Modern English, and since one of the most productive areas of development in the modern language is the ability of words to move across word classes, it may be preferable to accept what the corpus seems to be signalling, which is the need for a major overhaul of the notion of word class.

In general, we must move toward a theory that reconciles the paradigmatic and the syntagmatic dimensions and allows the description of the language to remain sensitive to both dimensions for as long as the correlation is productive; no doubt there will be some residue of specifically grammatical and specifically lexical information after that stage, but we must wait to see what it is, and what categories and processes are best used to describe it.

[1] (a) in Tickoo (ed.) 1989, reprinted as Chapter 6 of Sinclair 1991, I showed that the second commonest word in English, *of*, had very little in common with other prepositions, and was mainly used in a unique syntactic function.
(b) I followed this up in Sinclair (1999), where I argued that most of the common words in English have individual patterns of occurrence, and do not fit into the general word-classes.
(c) As a contribution to the NERC Report (1996), I pointed out that in English, as well as words which function as nouns, and those which function as verbs, there is a substantial class which function as both (I called them *norbs*). This is a kind of underspecification which postpones a very difficult set of decisions until perhaps the analytical system is better able to deal with them.

Lexical Structure

At present, the lexical structure is presented separately, insensitive to the grammar in the same way as the grammar is, traditionally, insensitive to the lexis. It is probably a valuable exercise to prioritise the lexical patterning and to push a lexical description as far as it is reasonable to do so; the justification is that so little research has been done in this area, especially as compared with the immense attention that the grammar has had over the centuries. But such an effort should not be misunderstood; it must be seen simply as an interim step towards an eventual holistic description, and there is no imperialistic dimension to lexical description.

In the meantime, there are structures of a particularly lexical nature that are worthy of attention, and which are introduced in recent publications, particularly Sinclair (1998). These begin with *collocation*, the co-occurrence of words, and go on to *colligation*, which in this work is defined as the co-occurrence of words with grammatical choices, then *semantic preference*, which is the co-occurrence of words with semantic choices, and *semantic prosody*. The semantic prosodies express attitudinal and pragmatic meaning; they are the junction of form and function. The reason why we choose to express ourselves in one way rather than another is coded in the prosody, which is an obligatory component of a lexical item.

The ways in which the prosody is expressed are extremely varied, and seem to have no limits as to position or shape; we can thus anticipate severe technical problems in retrieving them computationally. This is the central problem in analyzing open text and one of the principal reasons that the performance of devices which depend on some kind of language understanding is so poor. At the present time the goal of the machine understanding of language is far more difficult than it needs to be, because we are not using appropriate theories – once the meaning created by lexical structures becomes available, and integrated with what we already know through grammar, then theories will be articulated that predict the prosodies and the computer will then know where to look for them. These theories will be developed from the kinds of hypotheses that are taking shape in corpus-driven linguistics.

Example

Let me give as a conclusion an example of the kind of semantic prosody that I'm talking about. Consider the English word *effort*; it is a countable noun and so it has a plural, *efforts*. And one of the most notable collocates of *efforts* is the word *to,* which follows *efforts,* and which is the infinite marker. So essentially we are focussing on a structure which has a core of *efforts* plus an infinitive. In the Bank of English[1] in Birmingham, which is the reference corpus that I normally use, the one that lies behind the *Cobuild* publications, there are 9,617 instances of *efforts* followed by *to*. For the Figure, the computer has selected 21 by the simple expedient of picking the first one in text sequence, then dividing 9,616 by 20 (= 480 in round figures) and then selecting the 481st, 961st etc instance through the corpus.

If we examine these, it becomes fairly clear that we use this phrasing – we talk of "efforts to" do something – when they appear to be very unlikely to succeed, to be heading for failure, or already unsuccessful. In other words, the prosody that appears in almost every example is the speaker/writer's prejudgement of the efforts, that they are heading for failure. So when we are discussing the machine understanding of language, if we were to talk of the "efforts" of computational linguistics "to" comprehend natural language, we would imply that they are doomed to failure. There are a number of adjectives, for example, like *hysterical*, *frantic*, *futile*, *strenuous*; verbs like *blunder*, *hamper*,

[1] At the time of retrieval the Bank contained almost 350 million words of broad general English text, from native speakers in many parts of the world, their spoken and written expression. The corpus is jointly owned by HarperCollins, publishers, and The University of Birmingham, and access to it can be arranged via the Cobuild Home Page, <www.cobuild.collins.co.uk>

were overwhelmed; people *close ranks against* efforts, or achieve things *despite efforts*; efforts *exhaust* us, and so on. So if this is a representative sample of the behaviour of the word, we can expect to find in the left-hand cotext of *efforts to,* some indication of the likely failure of the efforts.

In a contrast which is almost ironic, we can expect to find in the right-hand cotext a set of verbs which are creative, which talk about creative action, like *please, revive, work together, protect, support, gain, raise, activate, kindle (a debate), help, give (the city something good), save,* etc. So before a reader/listener discovers that the efforts are to do something constructive and beneficial, they are already sabotaged. Our first draft of the lexical item that has as its core *efforts to* will thus contain three elements of structure – the core, the semantic preference for a verb of constructive action, and the semantic prosody of anticipated failure. The selection of the item is controlled by the prosody, because the whole point of expressing oneself in this way is to pre-evaluate the actions, which would otherwise be positively evaluated by the reader/listener.

It is likely that other expressions with structural similarity to this tentative item will be found; *attempts to* may be similar,[1] etc. The singular forms, *effort, attempt* may show some tendency in the same direction. A set of forms sharing similar meanings could be a further step in the mapping of an organisational framework for lexis. This work is only beginning; a few probes have been made into the lexical structure of the language, and some tentative hypotheses have been formulated.

[1] The appearance of *despite, failed, unsuccessful, desperate, repeated, several,* as very significant collocates of *attempts to* suggests considerable similarity, but measures have not yet been devised to compare collocational profiles.

Conclusion

Despite the recent rush to welcome corpora into the resource collections of many students of language, we must note that the vast majority of work with corpora still takes place under the assumptions of pre-corpus linguistics, and is thus insensitive to the possibilities put forward here. It is clear that the first step towards a new view of language has now been taken by the linguistics profession, in recognising that corpora are relevant and useful; this has been effectively completed approximately thirty years from the advent of electronic corpora. It is only natural that to begin with scholars will appreciate the security of familiar concepts in engaging with such a total revolution in the availability of evidence of usage, and only gradually will they accept that some of those concepts are sorely in need of being revised and updated.

The initial separation of grammar and lexis in language description, and the subsequent prioritisation of grammar at the expense of lexis, is one of the most firmly-held positions among theoretical and descriptive linguists, and it will take some time before it is held up to scrutiny and approached with an open mind.

References

Fan, M. 1999. "An investigation into the pervasiveness of delexical chunks in authentic language use and the problems they present to L2 language learners." In: Berry, R. et al. (eds.): *Language Analysis, Description and Pedagogy.* Language Centre, HKUST, Hong Kong.

Miller, G. 1956. "The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information." *The Psychological Review*, vol. 63:81–97.

NERC Report 1996; *Network of European Reference Corpora* Project Report. Pisa, Giardini.

Sinclair, J. 1991. *Corpus Concordance Collocation*. Oxford, OUP.

Sinclair, J. 1996. "The Search for Units of Meaning." In: *TEXTUS* vol. IX no. 1.

Sinclair, J. 1999a. "A Way with Common Words." In: Hasselgård, H. and S. Oksefjell: *Out of Corpora*. Rodopi, Amsterdam & Atlanta.

Sinclair, J. 1999b. "The Lexical Item." In: Weigand, E. (ed.): *Contrastive Lexical Semantics*. Amsterdam/Philadelphia: John Benjamins. (Volume 17 of series *Current Issues in Linguistic Theory*), pages 1–24.

Reference Works

Collins English Dictionary, Hanks et al. 1998.

Collins Cobuild English Language Dictionary, Sinclair et al. 1987.

Today's English Dictionary, Sinclair et al. 1995.