The title of a paper by Jim McCawley (nd) puts the main theme of my paper in a nutshell. His paper is called 'What linguists might contribute to dictionary-making if they could get their act together', and I shall argue that linguists in some of the better-known schools (such as are described in most introductory textbooks) haven't yet 'got their act together' on a number of questions which are of crucial concern to dictionary-makers. It is true that such linguists have some important and sophisticated things to say about some of the other parts of a lexicographer's work — I have in mind such matters as the nature of phonological representations, and some types of statement about valency (alias strict subcategorisation). However I imagine most lexicographers are already aware of these achievements of linguistics, so I shall have little to say about them. (In case a check-list of lexical knowledge is of interest I have included one as an appendix to the paper.) I think it is much more important to warn lexicographers against taking too seriously some very general claims of linguists that touch on their work.

The basic problem is that any of us linguists is also a citizen, with the same experiences as any other typical citizen. We all have dictionaries on our shelves, and have grown up in a society where dictionaries are standard items of furniture. These dictionaries are of course the traditional commercial ones, which have various structural characteristics. One is that they distinguish between 'the dictionary' and 'the grammar', the latter being either printed in summary as an appendix, or left out altogether. I think it is at least partly because of this institutionalised distinction that so many theoretical linguists are convinced that human language has a similar organisation: it consists of a set of rules plus a lexicon. Anyone who has done any detailed work on grammar knows that it is anything but obvious that there is a distinction between two radically different types of organisation, and it is at least a plausible hypothesis that linguists like to think there is such a distinction because it corresponds so
closely to the one institutionalised by commercial dictionaries. Similar points are very easy to make in relation to the internal organisation of this putative lexicon – e.g. the fact that the lexicon is generally assumed to consist of a list of discrete lexical entries could well be related to the fact that commercial dictionaries consist of a list of entries, each treated as a separate 'paragraph'.

What I am suggesting, then, is that folk linguistics contains various ideas about the lexicon, alias dictionary, which are at least in part founded on the traditional practice of lexicographers. Any linguist brought up in a culture where these folk ideas are prevalent is likely to be infected by them in early life and must beware of building them, without critical examination, into their professional thinking. And forward-looking lexicographers must be even more careful not to mistake ideas which originated in traditional lexicography for carefully considered and researched tenets of scientific linguistics. I applaud any attempt by a lexicographer to learn from linguistics (or from any other relevant discipline), in the hope of being able to move towards radically new and better kinds of dictionary; but it would be tragic if the effect of this contact with linguistics was just to tie lexicographers even more firmly to their own tradition.

Folk linguistics may of course be right on a particular point, but it may equally be wrong, and each claim must be examined in its own right. It could be argued that the very persistence of the commercial dictionary as an institution is evidence for the reality of the concept 'lexicon', and that the internal organisation of these dictionaries is in some sense natural because of a close correspondence to the way in which language is in fact organised in our minds. However there are other explanations for the success of traditional dictionaries – e.g. the lack of any alternative source for the same information, and the particular range of uses to which such dictionaries have been put – and it is at least conceivable that these explanations are sufficient in themselves; that dictionaries have been sold and used over the centuries in spite of a gross mismatch with the way language is 'really' organised. I shall argue below that this is in fact the case.

An important preliminary question arises about aims and evaluation criteria. What do we mean by the 'real' organisation of language? I recognise that linguists have traditionally fallen into two types, those who see language as a mental phenomenon and those who see it as in some sense 'out there', as a property of a society. However I have the strong impression that the mentalists have more or less taken over, and linguists would agree generally that language is at least a mental phenomenon – it is known by people – whatever else it may be. So it would be fair to say that linguists are aiming at psychological reality in some form or other; and when we argue (for instance) about whether the lexicon is separate from the rules of the grammar, we are arguing about the existence of some distinction in the structure of people's minds. Claims by linguists can thus be taken as claims about mental organisation.

The question then arises whether this kind of reality should be of interest to
lexicographers. After all, historically a good deal of lexicographic work has involved combing through written texts where questions of psychological reality appear somewhat remote. Moreover, any dictionary is inevitably to some extent prescriptive (if only regarding spellings), because ordinary users (or would-be users) of the language concerned consult it as an authority. This means that the dictionary takes on a kind of external objectivity that goes beyond the mental structures of any individual. If a dictionary is to be successful (and saleable) it must go beyond the knowledge of any given individual in scope – the whole point of having a dictionary in your house is to give you access to a wider range of linguistic knowledge than you have built up yourself on the basis of your own immediate experience (or to jog your memory, which in a sense amounts to the same thing). Nor is there any reason why a dictionary should limit its total capacity to whatever we think is the maximum available to a human – and I imagine that the OED must ‘know’ more words than any human could possibly know.

In spite of these observations which seem to suggest that psychological reality is irrelevant to lexicographers, there are reasons for thinking otherwise. One is that the information derived from traditional data collection always has in fact been supplemented heavily by the lexicographer’s interpretation of the words concerned – most obviously regarding their meanings, but also regarding their part of speech and so on. So a dictionary does represent one psychological reality willy nilly, namely that of the lexicographer. Another reason is that a dictionary is written so that its users will be able to learn what at least some people know about whichever words are looked up – let us call these people the ‘model speakers’. It is not enough to show the observable features of the words as used by model speakers – i.e. how they pronounce and spell them – but we also want to know what they mean by them, how they classify them grammatically, and so on. A third reason is of course particularly relevant to machine-readable dictionaries, which is that some people want dictionaries for use in AI research, where questions of psychological reality are paramount. And lastly there is the fact that the users of a dictionary are themselves human, so the information presented must be such that they can build it into their mental lexicon.

I shall take it for granted, then, that linguists and lexicographers are both in pursuit of the same goal, which is truth defined in terms of psychological reality. Arguments by linguists for and against particular views about the lexicon should be relevant to lexicographers – and of course the work of lexicographers is of great relevance to linguists because many of our hypotheses can only be tested with the help of a decent database produced by a lexicographer. Needless to say, accepting the goal of psychological reality does not guarantee that we shall achieve it, or know when we have achieved it, or even know precisely what it would mean to have achieved it. The point is simply that linguists do not make their findings irrelevant to lexicographers simply by setting themselves this target.
1.2 Main-stream linguistics

As we all know, there is no such thing as a single theory of linguistics to which all linguists subscribe — though there are a large number of individual points on which linguists seem to speak with one voice (in a recent survey of about 50 British colleagues I mustered no fewer than 83 such points — see Hudson, 1981). The theory of grammar, which includes the theory of the lexicon, is particularly prone to fission, and a good number of alternative theories for all or part of this area are available. In this paper I shall nevertheless refer to the notion ‘mainstream linguistics’ as a short-hand for a particular group of theories which I assume lexicographers and others are most likely to be aware of. I should make it clear that my own theory could not conceivably be counted as part of mainstream linguistics, but I am uncertain about a number of theories — for example, Systemic Grammar (Butler, 1985) is quite well known in computational circles, though much less so in theoretical linguistics. Inevitably some of my colleagues will be offended by my decisions, but I think the group of theories that I have identified as ‘mainstream’ have enough in common to make significant generalisations possible; and they are all certainly well known in the linguistics community.

The group consists of theories which are heavily influenced by Chomsky, including some which are explicitly opposed to Chomsky’s current position. They are:

a. EST — the Extended Standard Theory of transformational grammar (e.g. Radford, 1981);  
b. GB — Chomsky’s current theory, called the Government and Binding theory (e.g. Chomsky, 1981);  
c. GPSG — Generalised Phrase Structure Grammar (Gazdar et al. 1985)  
d. LFG — Lexical Functional Grammar (Bresnan, 1982).

1.3 Some questionable tenets of the main-stream

I shall now point out and comment on some of the more general tenets of mainstream linguistics which are relevant to lexicography. I shall return to these tenets in the second part of the paper, where I offer an alternative. To summarise my argument, I shall dispute all of the following widely accepted assumptions:

a. that the lexicon is a distinct component of the grammar;  
b. that there are discrete lexical entries;  
c. that the ‘lexico-grammar’ contains only intra-linguistic information.

In each case I shall call into question a boundary:

a. the one between the lexicon and the (rest of the) grammar;  
b. the one between one lexical entry and another;  
c. the one between language and other kinds of knowledge.
The general problem with these boundaries, I shall suggest, is that we have no general principles, with sound theoretical support, on which we can base decisions about where the boundaries lie in particular cases of uncertainty. In other words, my contention is that these boundaries are all absent from the kind of reality which we are trying to model, so if we include them in our theories we are inventing them rather than discovering them.

Linguists often say that such boundaries are essential if we want a linguistic theory to be as strong (i.e. as constraining) as possible – as we surely do. Removing the boundaries might make the linguist’s (and lexicographer’s) job easier, by side-stepping a range of difficult decisions, but it would so weaken our theories as to make them incapable of explaining anything. This view needs to be taken very seriously. We must be clear about what precisely is in question.

The general principle that theories should be as constraining as possible is beyond dispute. What I doubt is whether these particular boundaries actually help in achieving this goal. It is hard to think of any property of linguistic structure (or use, for that matter) which is explained by the presence of one of these boundaries – any proposition which has the form: ‘language has such-and-such property because there is a boundary between the lexicon and the grammar (or between lexical entries, or between language and non-linguistic knowledge)’. Some attempts have indeed been made to show that the boundary between the lexicon and the rest of the grammar accounts for various facts, but I shall suggest that these attempts have not yet been successful. If, then, the boundaries do not actually do any work in strengthening our theories about language, we may as well remove them from the theories and concentrate on developing all the other parts of the theories which do give them strength.

As far as lexicography is concerned, all three boundaries generally seem to be treated in an ad hoc and untheoretical way. If they are unreal, this is both right and inevitable – lexicographers have specific practical concerns which guide them in their decisions, and they neither need the help of theoretical linguists in making them nor (in this case) would they benefit from such help if it rested on the main-stream assumptions. If it turns out that lexicography is better if dictionaries are separated from grammars and encyclopedias, or if dictionaries are divided into discrete entries, then so be it; but if not, then lexicographers should have no theoretical compunctions in jettisoning the boundaries in question.

a. The lexicon is a distinct component of the grammar – the grammar (in a relatively general use of the word) consists of the rules (the ‘grammar proper’) and the lexicon. In order to avoid confusion about the meaning of ‘grammar’, I shall use the term lexico-grammar (borrowed from systemic grammar – Butler, 1985) to mean the combination of all the general rules plus the lexicon, and reserve ‘grammar’ for the rules; even in this narrower sense it is still ambiguous according to whether it includes e.g. phonological rules or not, but this does not matter for present purposes.
The idea that grammar and lexicon are distinct has been the standard wisdom since Chomsky 1965, though Chomsky’s earlier work did not recognise any such distinction. At one time it was fairly clear in what sense the lexicon was distinct: when generation was seen as a process which converted a single input symbol (S) into a full sentence-structure the individual operations had to apply in some order, and those operations which belonged to the lexicon had to apply before one particular stage in the generation process, namely ‘lexical insertion’. Nowadays, however, none of these theories accept this interpretation of a lexico-grammar. Rather, all the relevant bits apply simultaneously to check a ready-made structure for well-formedness, so the notion of applying ‘before’ or ‘after’ lexical insertion, and even lexical insertion itself, mean little.

Moreover, in the early days of the lexicon it was fairly clear that the lexicon was the repository for irregular and unpredictable properties of words (e.g. Chomsky, 1965: 87). However, it was soon suggested that redundancy rules and word-formation rules are needed (e.g. for derivational morphology and compounding) as part of a lexico-grammar; and since these seemed to create new lexical items it was assumed that they must be part of the lexicon, although this meant that the lexicon would then contain some highly regular patterns. The extended view of the lexicon raised serious problems which – to my mind – have hardly been addressed, let alone solved, in the two decades that have passed since they arose.

In particular, what is it about ‘lexical’ rules that distinguishes them from other rules, and that makes them suitable for inclusion in the lexicon? The best attempt to answer this question is Wasow 1977, which suggests a set of five tests for distinguishing lexical rules from transformations, but these are at best a reflection of then-current practice and hardly a justification for this practice. The kind of question that arises is illustrated by the rule for creating agentive nouns by adding -er to a verb. Is this a lexical rule, because it relates words of different major classes, or is it a transformation – or more generally a non-lexical rule – on the grounds that it is relatively productive and regular? In each case the justification suggested is in Wasow’s list of tests, and there is no general principle to which we can turn in order to resolve such conflicts.

A related set of principles is given by Kaplan and Bresnan (1982: 180): ‘Rules that change relations are lexical and range over finite sets, while syntactic rules that project onto an infinite set of sentences preserve grammatical relations.’ This suggests the basis for a real distinction between rules that belong in the lexicon and those that don’t. Of course it cannot be the case that any rule which leaves grammatical relations unchanged is outside the lexicon, because this is true of many standard word-formation rules (e.g. the one for prefixing un- to adjectives), and Bresnan herself takes all inflectional rules as lexical (Bresnan, 1982: 18); so we still lack a way of telling whether a rule which preserves grammatical relations is lexical or not. The principles given by Kaplan might however work in the opposite direction: it might be possible to conclude that any rule which changes grammatical relations has whatever other properties
one would expect of lexical rules, such as that of referring only to material which is available in lexical entries—i.e. subjects and complements, but not adjuncts or other modifiers (Kaplan and Bresnan, 1982a: 214). There are reasons for doubting whether this claim is true.

Take a passive sentence like the attested *This paper has been written on both sides of.* Passivation is clearly a rule which changes grammatical relations, and is one of the main examples in LFG of a lexical rule. Prepositional passives are handled by means of an ‘incorporation’ rule (Bresnan, 1982b: 52) which reanalyses a preposition after the verb as part of a complex verb, thereby changing the preposition’s complement into the complement of the verb. However, this will only work if the material to be incorporated is already present in the verb’s lexical entry, and it seems unlikely that this is the case for an example as complex as the one quoted. (Note that *both sides* could have been *one side* or *-, if paper had numerous sides — thirteen sides.*) Similar remarks apply to many other examples of prepositional passives, like *This bridge could never have been flown under by a plane like that* (Davidson, 1980). It seems unrealistic to assume that the verb *fly* is subcategorised for the preposition *under.* One could even generalise to other constructions, such as the indirect objects related to *for* phrases—a clear case for a relation-changing rule. For example, it is hard to believe that the entry for a verb like *peel* refers to the beneficiary (as in *Peel me a grape!*).

These examples are meant to suggest that some rules may combine properties of lexical rules (e.g. they change grammatical relations) with those of non-lexical rules (e.g. they refer to material which is absent from lexical entries). And yet according to the standard assumptions any rule must be classified as either lexical or non-lexical, so what are we to do about such cases? At the very least we can see that the boundary between the lexical and the non-lexical requires very careful scrutiny indeed, to make sure it really is there. However there may well be strong tendencies for rule-properties to occur in clusters—e.g. for relation-changing rules also to refer to subcategorised material—and it would be a pity to throw out the baby with the bath-water when abandoning the distinction between the lexicon and the grammar. The notion ‘lexical rule’ looks like a prototype, organised around clear cases but with deviation permitted in its instances. Indeed, I have argued that all concepts in linguistics are prototypes (Hudson, 1980: 233f, 1984: 39ff, 1986a). But this is not how main-stream linguistic theories (including LFG) present the lexicon: any given item of information is assumed to be either inside the lexicon, or outside it—‘degree of lexicality’ is a notion which is conspicuously absent from the discussions.

The problem of deciding which rules are lexical and which are not is not the only one that faces proponents of the lexicon-grammar boundary. Another is that of deciding what to do with examples like *What about a drink?* and *Down with the government!*—which are ‘non-canonical constructions’. The problem is that these sentences have structures which are ‘non-canonical’ in the sense
of not fitting into the normal patterns – e.g. What about a drink? consists of nothing but a question word plus a prepositional phrase, and there is no hint of a fuller abstract structure from which this could have been derived. So we need a rule of some kind to generate the structural pattern found in each such case; but this rule is tied to just one or two lexical items (e.g. what can be replaced only by how). Thus on the one hand the pattern looks lexical, because it belongs to just one or two lexical items; but on the other hand it looks grammatical because it alone is responsible for the structure in which these lexical items can occur. Once again there seems to be no gain, and considerable loss, if we are forced to decide categorically between the lexicon and the grammar as the place to locate the information in question.

In short, a much more serious appraisal is needed of the notion 'lexicon'. All that has happened since the notion 'lexicon' was introduced is that various unprincipled decisions have been made, have been accepted, and have turned into standard practice. For example inflectional morphology, however regular, is now dealt with by some (e.g. Bresnan, as mentioned above) in lexical rules. There is now even a movement called 'lexical grammar' (represented e.g. by the papers in Hoekstra et al. 1980 and Moortgat et al. 1981), whose driving force is the belief that a good deal of lexico-grammar should be handled 'in the lexicon' – though there is virtually no attempt to explain precisely what this belief means. It is hard to avoid the conclusion that this whole area of activity is concerned with nothing but terminology.

How does all this affect the lexicographer? First, I am sure that any lexicographer must agree with at least the spirit of what I have said about the distinction between the lexicon and the grammar. In practice they must often be faced with unanswerable questions about what information to include in their dictionary and what to leave out on the grounds that it belongs rightly in a grammar.

For example, what kinds of word-order facts about English should go into a dictionary? Take enough: this is exceptional in following its head (compare sufficiently big with *enough big, and big enough with *big sufficiently). Presumably this fact should be included in the dictionary. Now consider in contrast the fact that the subject of a verb normally precedes it; this obviously belongs in the grammar, not in the dictionary. But what about facts intermediate in generality between these two examples? For example, indefinite pronouns (a fairly large, but closed, class containing words like someone and nothing) require their modifying adjective to follow them, contrary to the general pattern (someone nice, not *nice someone); auxiliary verbs (an even more limited class) allow their subjects to follow them in certain constructions, such as questions; and a very small list of auxiliary verbs can have a following subject and be used to introduce a conditional clause (e.g. Had I known that, ...). Which of these facts, if any, belong in the grammar? Obviously it is easy to make a decision, if only on the basis of flipping a coin; what we are short of is general principles for justifying the decision.
I see, and applaud, a tendency for commercial dictionaries to include a good deal of grammatical information (e.g. Longman's Dictionary of Contemporary English); and of course commercial grammars (e.g. Quirk et al. 1972) contain a vast amount of 'lexical information' in the sense of information about particular words. I take it that this trend shows that lexicographers are in fact cocking a healthy snook at main-stream linguists. However, if there is in fact no natural boundary between the lexicon and the grammar, then professional linguists should develop theories which reflect this fact. Such theories are bound to be of more interest to lexicographers than the current main-stream ones.

In conclusion, then, I have suggested that main-stream linguistics is short of theoretical underpinnings for the assumption that a lexico-grammar contains a distinct lexicon; and I have hinted that the reason for this shortage is that there is no distinct lexicon in nature - i.e. in our minds. If this is so consequences are bound to follow for the practice of both grammarians and lexicographers, whose respective patches turn out to be like the north end and the south end of the same field, rather than like different fields separated by a hedge. For example, if a particular fact could equally well be treated by either of them, it ought to be handled by means of the same formal system by both - and not, for example, as part of a static bundle of properties which cut across linguistic levels by the lexicographer, and by an if-then rule unrelated to rules about other levels by the grammarian. Main-stream linguistic theories all offer quite different formal systems for the lexicon and for the rules of grammar, so in this respect they are unsuccessful.

b. The lexicon is a list of discrete lexical entries. The notion of a 'lexical entry' is taken for granted by all the mainstream theorists. They all agree in seeing it as some kind of package of information from various linguistic levels (typically phonological, syntactic and semantic information, with or without morphological information). Take the following discussion from Gazdar et al. (1985: 34), for example:

.. let us assume that a lexical entry contains at least four kinds of information: a phonological form, a [syntactic] category, an indication of any irregular morphology, and a meaning. Keeping to orthographical representations for convenience, we might expect entries of the following kind:

\[
\text{\textit{weep},}
\]

\[
\text{\{\text{[I-N], [+V], [BAR 0], [SUBCAT 1]\},}
\]

\[
\text{\{wept\},}
\]

\[
\text{\textit{weep'}}
\]

A somewhat more complex, but perhaps more explicit, notation is used for lexical entries in LFG, but there does not seem to be any standard GB notation - indeed, it is rare to find examples of lexical entries in the GB literature.

The point of a lexical entry is to show what information belongs together; thus in the GPSG example just given the information between the diamond
brackets all helps to define the same lexical item (i.e. one whose spelling is *weep*, whose syntax is that of an intransitive verb, whose past tense is irregularly *wept*, and whose meaning is the concept named *weep*). The boundary of the entry, marked by the diamond brackets, is thus crucial, because there is nothing else which shows that the syntactic facts refer to the verb *weep* rather than to some other word. The same is of course true of the entries in a traditional dictionary, where the equivalent of the GPSG diamond brackets is the line-break between entries; and one suspects that the similarity is not coincidental.

Once you make the notion ‘lexical entry’ this important, you create problems (which are not dissimilar from those which arise when you assume a distinct lexicon) – all those analytical problems about boundaries between entries. Assuming that some disjunctions are allowed within an entry, how different may the alternatives within a single entry be? For instance, take the problem of distinguishing between polysemy and homonymy; in polysemy there is a single entry with more than one meaning (e.g. the sex-neutral and female meanings of *cow*), but in homonymy there are distinct entries (e.g. the ‘harbour’ and ‘drink’ meanings of *port*). As Lyons argues, there is no satisfactory way to solve this problem without effectively abandoning the distinction between polysemy and homonymy.

The polysemy/homonymy distinction is probably the most familiar problem connected with boundaries between lexical entries, but there are others too which arise when the alternatives are not meanings but forms. How similar must two forms be to count as part of the same lexical entry? Inflected forms are allowed in, as forms of the same lexical item, or ‘lexeme’; and this even when they are suppletive (e.g. *went* is accepted as part of the entry for *go*). Morphologically derived forms may or may not be included – there has been some debate about this in the linguistics literature, but the result seems at best inconclusive because of the lack of clear principles on which to base a discussion. Then there are lexically specified alternative forms which don’t seem to fit into derivational patterns – e.g. full and abbreviated forms of English auxiliary verbs (is versus ‘s, and so on); how similar do these have to be to belong to the same lexical entry? What about pairs like *I* and *me*? Linguists have always been reluctant to take such different forms as part of the same lexical entry, but is this for any reason other than because this would be a departure from the model of traditional lexicography? Once again there is no general principle to which one can point as a guide to making these decisions, which suggests that reality – i.e. mental reality – is not in fact structured like that.

A further problem with this reliance on lexical entries is that it makes no explicit connections at all among lexical entries. Indeed it is often assumed that to put two words in different lexical entries is to deny any connection at all between them, so that lexical relatedness is an all-or-none matter: either two words are related, in which case they will be shown in the same entry; or they are not related, and are in distinct entries. It seems unlikely, to say the least, that this is how our minds are structured; surely some words are more closely,
or more obviously, related than others are, but there is no natural break in the scale of relatedness. Take the verb *stand*, 'be upright'. This is closely related to its participle *standing*, slightly less closely related to the past tense *stood*, and more distantly related to *stand*, 'put up with', to the noun *stand*, 'fixed public decision' (e.g. *take a stand on some issue*) and 'raised stage' and 'place for showing things'. It seems meaningless to ask at what point we switch from sameness to total difference. In any case, even those who put such trust in the boundaries between entries also invoke other means of relating words to each other – e.g. synonyms, antonyms etc. can presumably be related without invoking lexical entry boundaries; and of course one of the functions of the lexical rules mentioned earlier is precisely to make connections among lexical entries. So it is unclear why the boundaries between entries need have any analytical significance at all – in which case, the question arises why we should bother chopping the lexicon up into entries at all.

It looks, then, as though the status of lexical entries as discrete units is in doubt. Let us now turn to the other side of the belief currently under review, namely that the lexicon is a list of lexical entries. What is the structure of this list? Main-stream linguists generally say it is just a set, or more precisely an unordered set (Chomsky, 1965: 84). Psycholinguists however argue that there must be ordering of some sort – for example, they argue that slips of the tongue show interference from lexical items that are 'near' to the target in the lexicon. If it could be shown that a single ordering is imposed on the entries in the lexicon, the difference between these two positions wouldn't matter – linguists leave the question open, and psycholinguists answer it. However there are reasons for thinking that entries are ordered on more than one dimension – a phonological ordering shows phonological relatedness (e.g. Fay and Cutler, 1977), a semantic ordering shows semantic relatedness (e.g. Lesser, 1978: 72), and so on.

If we could trust the notion of 'lexical entry', we might consider a single ordering of entries on the basis of some single factor – such as frequency or recency of use – and then add on various access systems, each based on a different type of similarity (e.g. Matthei and Roeper, 1983: 188). This would then give yet another way of showing relatedness among words, which would make the 'entry' concept even more redundant.

Alternatively, we can abandon the entry as the organisational unit of the lexicon. The units with which we are left are:

- whatever word forms are stored (I assume that regularly inflected forms are generally not stored, nor are variant pronunciations due to regular sandhi processes);
- whatever word meanings are stored.

The range of word forms and word meanings which need to be referred to in this approach may be precisely the same as those which are used in an entry-based approach, but instead of relating the forms and meanings by enclosing
them in a common entry, we relate them directly: 'word X means Y', and so on. This allows us to show one-many connections in either direction, without worrying about whether they cross entry boundaries; for example, we can include all the following statements:

- *stand* means ‘be upright’
- *stand* means ‘tolerate’
- *tolerate* means ‘tolerate’

These statements show the facts clearly without requiring us to agonise about whether the two meanings of *stand* are part of the same lexical entry or not. (Notice that agonising would be in order because the morphology is irregular in precisely the same way whichever of these very different meanings the word has.)

Of course, some word-forms must be related to each other in the lexicon, and if we have no entries to do this for us we must find an alternative. However, the entry is a very crude way indeed of showing such relatedness – consider, for example, the GPSG entry for *weep* which I quoted above, where the word *wept* is simply included in the entry without any explanation of its relation to *weep*. What we need for such cases is a precise statement such as ‘the past tense form of *weep* is *wept*’, or in the case of *stand*, ‘the past tense of *stand* is *stood*’. Once this information is available, in this form, there is no more to be said (and the content of the lexico-grammar would be precisely the same whether we included an entry boundary between *stand* and *stood* or not).

There is no need to assume in this system that words whose pronunciation or spelling is the same must have the same name in the lexicon. This is so because further statements are needed in any case to relate the words to their pronunciations and these statements can introduce further one:many relations – for example, we can give different names to the verb and the noun *stand* – say, ‘*stand*/verb’ and ‘*stand*/noun’ – and show the similarity between them via the statements about their pronunciations (just as we showed the synonymy of *stand* and *tolerate* via the statements about their meanings). Nor need we assume that syntactic differences have to be ignored; for example, we can distinguish ‘*stand*/verb’ from more particular instances of it, namely ‘*stand*/transitive’ and ‘*stand*/intransitive’. All three concepts would be needed in the lexicon: the first for the spelling, phonology and morphology, and the second and third for the syntax and semantics. The part of the lexicon which deals with these relations would thus contain statements such as the following:

- *stand* is spelled *stand*
- the past-tense of *stand* is *stood*
- *stood* is spelled *stood*
- *stand*/transitive is an instance of *stand*
- *stand*/transitive means ‘tolerate’
- *tolerate* means ‘tolerate’
The Linguistic Foundations for Lexical Research and Dictionary Design 299

- *stand* intransitive is an instance of *stand*
- *stand* intransitive means ‘be upright’

I shall develop these ideas in the second part of the paper, but what I hope to have shown is that the notion ‘lexical entry’ does not deserve the central place it is given in the main-stream. It is much too crude and inflexible an instrument for showing all the relations that need to be shown, and if other devices are used for showing these relations it no longer has any work to do. What we have moved towards, I think, is a conception of the lexicon much more like a network or relational data-base. This may not be much comfort for lexicographers who have to produce dictionaries in the form of books, but it is highly relevant for those who are working towards computerised dictionary data-bases which can be accessed in flexible ways. If the relevant field is meaning, then the transitive *stand* is grouped with *tolerate*, but if it is morphology, then the transitive and intransitive *stand* should be treated together.

c. *The lexicon contains only intra-linguistic information.* My last grumble about the main-stream is about the range of information which is given in the lexicon (and, more generally, in the lexico-grammar). It is assumed, without discussion, that the lexicon deals with nothing but phonological, syntactic and semantic information – and when an author happens to remember it, morphological information. (Once again the quotation from Gazdar et al. (1985) in connection with the entry for *weep* is typical, except that it allows for the possibility that other things might be included.) The list includes all and only those bits of information which refer exclusively to linguistic structures – i.e. to ‘strictly linguistic’ structures, structures which are part of ‘language’ and not part of – say – the encyclopedic knowledge we have in our minds. We can ask two questions about this approach. First, why should non-linguistic structures not be referred to in the lexicon? And second, how clear and important is the distinction between language and other kinds of knowledge? We shall see that a number of very general, almost ideological, assumptions are at stake.

Let us take an example of the kind of knowledge which is excluded because it involves non-linguistic structures: contextual restrictions to do with the kind of person who is speaking. For example, I know that *sidewalk* is used only by Americans, and I should expect a British dictionary of English to include this information. However, it cannot be included in a lexicon of the main-stream kind because the category ‘American’ is not a linguistic category, and only linguistic categories are admitted to the lexicon. Indeed, even the category ‘speaker’ is not a linguistic category, because it is not part of the analysis of a linguistic expression in terms of the standard list of linguistic levels (phonology, morphology, syntax or semantics); but why should this matter? If we say that the lexicon contains everything we know about lexical items, and if we accept the fact just quoted as part of most people’s knowledge about the lexical item *sidewalk*, then surely this should be included in the lexicon.

However, as I have already said, facts like this are not in fact included in the
lexicon – or anywhere else in the lexico-grammar for that matter – by main-stream linguists.

The principle behind this kind of exclusion is, I think, the very general belief in the separateness of language which has dominated linguistics throughout this century – one of the basic tenets of structural linguistics, in fact. The principle is often stated – indeed, it is often presented as a finding of scientific linguistics (e.g. Sperber and Wilson, 1986: 8), rather than as an initial premise – but there are no standard arguments for the separateness of language which those of us who deny it have to counter. Nor is there ever any attempt to explain why the principle should lead us to exclude all references to non-linguistic categories; after all, it would be quite consistent to believe that there is a clearly definable set of 'linguistic facts', but that what made them 'linguistic' was that each fact referred to some linguistic category – not that it referred to nothing except linguistic categories. The fact about sidewalk being used by Americans would then qualify as a linguistic fact by virtue of referring to a word, in spite of the fact that it also refers to a non-linguistic category. My conclusion on this question, then, is that there is no reason at all why reference to non-linguistic categories should be excluded from the lexicon.

Let me now give some other examples of the kinds of information which are not allowed to be given in a main-stream lexicon:

a. social constraints in general, including restrictions on type of speaker (e.g. our earlier example of sidewalk), type of addressee (e.g. gee-gee), formality of social situation (e.g. attempt versus try), etc.

b. encyclopedic information about the referent of a word (e.g. the definition of mouse given in the Longman Dictionary of Contemporary English: '... any of several types of small furry animal with a long tail, rather like a rat, that lives in houses and in fields ... – see picture at MAMMAL').

c. etymology – on the grounds that it refers to non-linguistic categories like language names and dates, not to mention the fact that etymological knowledge is extremely unevenly distributed through the population. It is worth remembering that in many communities most people are bilingual and therefore know something about the etymology of at least those recent loan words into one of their languages which come from their other language.

Nor is there, strictly speaking, any place even for spelling since many linguists believe that phonology is part of language structure but spelling is not. Whatever position one takes on this boundary dispute, the fact remains that spelling interacts with phonology on occasions (e.g. in spelling pronunciations), and that for some words some of us know a spelling but not a pronunciation. All of these exclusions, then, would be debatable even if one accepted the existence of a clear distinction between linguistic and non-linguistic categories, because the excluded statements are all about linguistic categories, and are all part of the knowledge of at least some speakers of the language.
The second question, however, is whether the distinction between linguistic and non-linguistic categories is itself either clear, or important. Let's assume that anything which we could classify as a word represents a linguistic category (while recognising that other kinds of linguistic category exist, of course – notably speech sounds). How clear is the boundary around the category 'word'? For example, what about *hey!*, or the filler *ummm*, or *shshl* (meaning 'be quiet')? Most people are unable to make any decisions about examples like these, even when they have no difficulty at all in using them – which seems to suggest that the distinction between words and non-words is not only unclear, but also not very important. These are very simple examples, but we could make a similar point in relation to semantic categories, which at least for GPSG and LFG are taken as part of the linguistic structure: how should we distinguish between categories which are semantic and those which are 'encyclopedic' (and therefore non-linguistic)? Once again the distinction seems to be of much greater interest and importance to linguists than it is to ordinary speakers.

My conclusion regarding the scope of the lexicon, then, is that there is no natural boundary around the types of information which are linguistic and therefore part of the lexicon and those which are not. In the previous two grumbles about mainstream linguistics I suggested that linguists had been too much influenced by traditional lexicographic practice in being seduced into believing in the 'lexicon' and the 'lexical entry'. Similar remarks could be made about the distinction between 'semantic' and 'encyclopedic' information – if it hadn't been for the existence of encyclopedias as well as dictionaries maybe the distinction wouldn't have suggested itself – but on the whole the problem in the present section is the reverse, namely that linguists have excluded from the lexicon so much information which has traditionally been included in dictionaries. If we assume that lexicographers will continue this practice, then it seems to me that forward-looking lexicographers have very little indeed to learn from mainstream linguistics, which in the three respects which I have discussed above lags well behind their practice.

2. Word Grammar

2.1 Overview

I now turn to the more positive part of my paper, where I present a theory which I think should be of more help to lexicographers than the mainstream theories discussed in the first part. The theory is called 'Word Grammar' (WG) because the central unit for linguistic analysis is the word; indeed, as I shall explain below, the lexico-grammar consists of nothing but statements about words – notions like 'phrase', 'clause' and 'sentence' play no part. In other ways too the theory is a radical departure from the tradition represented by the mainstream theories; for instance, it emphasises the continuities between the lexicon and the grammar, between the lexico-grammar and encyclopedic knowledge, and between the lexico-grammar and contextual knowledge. (Except where the
relation between 'grammar' and 'lexicon' is under discussion I shall use the term 'grammar' to mean 'lexico-grammar' as this is standard WG practice.) I shall do my best to explain the relevant parts of the theory below, but a brief overview may be helpful. The background assumptions and an earlier version of the theory are described in some detail in Hudson 1984, but the version which I shall assume here is rather different (see Hudson, 1985a, b, c, 1986 a, b, c, 1987, 1988, forthcoming a, b, Hudson and Van Langendonck, forthcoming).

In the more familiar linguistic theories, the structure of a sentence is described as a hierarchical arrangement of clauses, phrases and words – a 'constituent structure'. This is a rather un-traditional view of grammatical structure, although it is now so familiar that most linguists take it completely for granted. The traditional view, which is also the WG view, is that the structure of a sentence is best defined in terms of the relations between individual words – a 'dependency structure', because the relations concerned are asymmetrical dependency relations. For instance, take the sentence *Old cheese smells.* According to WG, the subject is *cheese,* which in turn has a modifier, *old;* but according to most other theories, the subject is the phrase *old cheese.* Similarly, in *Depend on me!*, the WG analysis recognises a direct link between *depend* and *on,* and another between *on* and *me.* In contrast, the constituent structure has *on me* as a phrase, and this whole phrase is related to the rest of the sentence.

As might be expected, the two theories also differ with regard to the other term of the relation. According to WG, and dependency theories in general, the other term is a word, so in *old cheese smells,* the word *cheese* is the subject of the verb *smells,* and *on* in *depend on me* is related directly to the verb *depend.* According to the constituent-structure analysis, on the other hand, the other term of the relation is a larger constituent, namely the whole clause or (for some analyses) a 'verb phrase'.

To summarise, then, dependency analysis is based on the relations between individual words, whereas the basis of constituent-structure analysis is the part:whole relation between a word or phrase and a larger phrase.

One of the technical problems with the constituent-structure approach is that many of the relations which need to be defined are indeed relations between words, and not between phrases. The example *Depend on me!* illustrates this problem nicely. As any lexicographer of English knows, there is a special relationship between verbs and prepositions: particular verbs select particular prepositions, in an often arbitrary and unpredictable way. A clear case is the verb *depend,* which selects *on.* Now if *on* is part of a phrase, there is no direct grammatical relation between *depend* and *on:* instead, there is a relation between *depend* and the whole phrase, and then also a relation between the phrase and one of its parts, *on.* In a dependency analysis, however, there is a direct relationship: *on* depends on the verb *depend,* and in our example *me* depends in turn on the word *on.*

The dependency analyses of WG also offer a vocabulary of grammatical relations of the kind that lexicographers need. Terms like 'subject' and 'object' are familiar from our grammatical tradition, but they are also found in WG
grammar, along with a range of somewhat less familiar terms - e.g. adjunct, complement, incomplement, oblique. The list of terms is quite short - about a dozen or so distinct relational categories seem to suffice for defining all the grammatical relations that need to be referred to in a grammar. Moreover they are arranged in a hierarchy, with the most general term, 'dependent', at the top, and increasingly specific terms lower down (e.g. a dependent may be a postdependent, which may be a complement, which may be an object, which may be a direct object). The most delicate distinctions are needed only for those words which allow the most dependents with distinguishable properties, and for other words far fewer distinctions are needed - thus verbs require the full set of distinctions, but prepositions allow very few of them.

The hierarchical arrangement of grammatical relations is of a type called an 'isa' hierarchy, for obvious reasons - each element in it is in the 'isa' relation to the element above it (a direct object isa object, an object isa complement, and so on). Isa hierarchies are fundamental to WG, and explain why there is no natural division between the grammar 'proper' and the lexicon. To see this, we must see how word-classes are handled. A hierarchy of 'word-types' is recognised, at the top of which we find the most general type of all, 'word', and at the bottom are all the individual lexical items (CAT, DOG, etc.). In between are some very general types like 'noun', but also some much more specific ones, like 'personal pronoun' or even 'reciprocal pronoun' (with just two synonymous sub-cases, each other and one another). As we have already seen, there are serious difficulties in deciding where to locate the boundary between 'grammar' and 'lexicon', but if these two areas are just 'the top' and 'the bottom' of a single hierarchy, that is hardly surprising.

The information contained in a WG grammar takes the form of a large number of simple propositions (none of which has any conditions, footnotes, diacritics or other paraphernalia). The notation is rather similar to ordinary English, which makes them relatively easy to read and to write, compared with the rather arcane notations of some other theories. Here are some examples:

[1] stem of eat is eat
[2] whole of past-tense of eat is ate
[3] eat has ano object
  [Note: 'ano...' = 'a or no...' i.e. 'an optional...']
[4] eat isa verb
[5] affected of referent of eat is referent of object of eat
[6] whole of past-tense of verb is stem + ED
[7] verb has ano subject
[8] tensed verb has a subject
[9] object isa postdependent
[10] subject isa predependent
[11] word precedes postdependent of word
[12] word follows predependent of word.
Examples [1] through [5] define some of the distinctive facts about the verb *eat*, while those in [6] though [12] are more general. It will be seen that precisely the same kind of notation is used for the general facts as for the specific, 'lexical' ones; indeed, in some cases the only difference is one of generality, as can be seen from a comparison of pairs like [3] and [7]. Some pairs of propositions contradict one another – for example according to [2], the past tense of *eat* is ate, but according to [6] it ought to be eat + ED, assuming that *eat* is a verb. In such cases the more specific proposition overrides the more general one. Lexicographers would all agree, I take it, that the first five facts come within their purview, but some more modern dictionaries include information as general as that in the remaining propositions. The similarities between the two kinds of fact, which are highlighted by the WG notation, help to explain why it may be easier to cover everything than to try to draw a line between lexical and grammatical facts.

2.2 Word grammar compared with the main-stream

I shall now take each of the three ‘questionable tenets’ of main-stream linguistics which I discussed in part 1 and explain what WG has to offer as an alternative.

a. The lexicon is not a distinct component of the lexico-grammar. As we have just seen, a WG grammar consists of a set of propositions, some of which refer to lexical items. We may want to refer to these propositions collectively as 'the lexicon' when talking about the grammar, but it is hard to see why one should do so, because of lack of important differences between such propositions and more general ones. It is true that it is only at or near the bottom of the hierarchy that one finds propositions about particular spellings or pronunciations (such as [1]), but this hardly seems to justify a major boundary.

One of the consequences of this hierarchical view of grammar is that we may expect the hierarchy to continue downwards, even below the lexical items. Why should it stop after we have reached elements such as *eat*? Why shouldn’t we recognise distinct sub-cases of *eat*, such as the *eat* which occurs with *humble pie* as its object? (More precisely, it would be the *eat* whose object is the sub-case of *pie* whose modifier is *humble*.) If we called this *eat/23*, or even *eat/pie*, then we would define its meaning in a proposition which would override the one for its superordinate, *eat*, following the usual principle that conflicts are always resolved in favour of the more particular case. Cutting a number of theoretical and descriptive corners we could represent the two meanings as follows:

[13] referent of *eat* is eating.
[14] referent of *eat/pie* is showing-humility.

The special syntactic requirements of the idiomatic use would be shown by a proposition that makes an object obligatory, contrary to the normal pattern of [3], and by another which requires it specifically to be the word *pie/humble*.
eat/pie has an object.

Object of eat/pie is a pie/humble.

Alternatively, it could be argued that there is no definable grammatical relation between eat and humble pie, beyond the simple fact that the latter immediately follows it. In this case we could simplify the whole analysis:

eat/pie has no object.

next of eat/pie is humble pie.

Many details remain to be worked out, but the example should suggest roughly how idioms might be analysed with the help of the isa hierarchy of word-types, on the assumption that it is open-ended (at the bottom).

According to WG, then, there is no point in asking where the boundary between the dictionary and the grammar ‘really’ lies; there is no such boundary in our minds, so it is only a practical convenience if publishers invent one. Indeed, we can go further: if publishers or computational linguists decided that it would be helpful to produce a single unified lexico-grammar, WG would provide theoretical justification and support. Equally, it would be just as easy to justify some way of dividing the data into manageable portions which had nothing to do with the division between the lexicon and the rules – for example, we could classify propositions according to whether their arguments referred to meaning or to form (or to both).

Nor is there any theoretical justification for fundamental differences in the operations that a parser or generator carries out according to whether it is accessing lexical or ‘grammatical’ information – e.g. a difference between ‘lexical look-up’ and some kind of tree-building operation. Once some stored word has been found as a potential model for an input word, a uniform process exploits all the knowledge which is known about that stored word, and also about its model in the grammar, and then about the latter’s model, and so on recursively. Once all these word-length models have been exploited there is nothing left to do.

If WG propositions can be used both for ‘lexical’ facts and for more general facts of grammar, what else can they be used for? One of the main theoretical claims of WG is that linguistic knowledge is just one particular case of knowledge, and that there are no formal differences between the structures found within language and those found outside it. To the extent that this hypothesis is valid we should be able to use the same formalism for the encyclopedic knowledge as for linguistic knowledge – which again is as it ought to be, given the well known difficulties of drawing the line between the two (e.g. Langacker, 1985; Haiman, 1980). The tendency to incorporate more and more ‘encyclopedic’ information into commercial dictionaries supports the idea that there is no natural distinction, and any theory which invokes different formal structures for the two kinds of knowledge should be suspect.

b. The lexicon is not a list of discrete lexical entries. In a WG grammar there are no packages of information which one could reasonably call ‘lexical
entries'. As we have seen the grammar is a list – or better, a network – of propositions, each of which has two arguments. A rough equivalent of a lexical entry is the set of propositions which refer to some given word W, but the equivalence is only very approximate. For one thing, these propositions are not picked out in any other way (e.g. by being assigned to a mini-component of the grammar called 'the lexical entry for W'). For another, a single proposition could refer to two different lexical items, and in such cases it would be arbitrary to assign the fact concerned to the lexical entry of one of the words rather than the other. For example, [16] relates the word eat/pie to the word pie/humble, so which lexical entry does it belong to – the one for the former, or the one for the latter?

Another difference between the network structure of WG and a structure based on lexical entries is in relation to examples like no and none (or my and mine, etc). Are we dealing here with two different lexical items, and therefore two lexical entries, or with just one? If possible it would be good to avoid making this decision, and this is possible in WG. I assume that the words concerned both have precisely the same meaning and syntactic constraints, except that the former must occur with a following common noun, and the latter must occur without one (hence No/*none time was left, but None/*no was left). I assume further that neither is more fundamental than the other. We can then postulate a more general 'lexical item', no(ne), of which each is an instance. The meaning and any other facts are stated in relation to this item, but it has no spelling/pronunciation, this being related to the more particular subcases, each of which also has a different syntactic restriction. Once again it is necessary to simplify the analysis and to beg a number of important questions for the sake of the example.

[19] quantity of referent of no(ne) is zero.
[21] whole of none is none.
[22] none has no complement.

In general, then, WG shows whether and how words are related to each other by means of propositions, in contrast with theories which invoke the rather crude mechanism of assigning words to lexical entries.

c. The lexicon does not contain only intra-linguistic information. Now that I have explained how WG presents linguistic knowledge as part of a much more comprehensive knowledge-structure, it is easy to see that a proposition could easily refer both to some word (or word-type) and also to some non-linguistic concept. The easiest example of this involves the isa hierarchy of word-types itself, as the concept 'word' is itself a particular case of the more general concept 'communicative action', and so on. Far from being a clearly bounded and uniquely structured 'module', as is sometimes claimed by linguists, language is just one area of a much larger network of knowledge, with very few distinctive characteristics other than its specifically linguistic content.
When one concept is an instance of another, the former 'inherits' the properties of the latter. We have already seen this in the linguistic hierarchies considered earlier – e.g. eat inherits the property of having an optional subject from a proposition which refers just to 'verb'. Now if we locate 'word' in the general hierarchy of concepts such as 'action', it follows that all the properties of these higher concepts, however 'non-linguistic', become available to 'word' (and via 'word', to all particular words and word-types). One of the properties of the general notion 'action' is that it has an actor (in contrast with other kinds of event, which just happen and have no actor). So words too have actors, namely, their speakers. This rather unsurprising observation is in fact incompatible with every other theory of grammar, since these are explicitly and deliberately designed only to accommodate intralinguistic facts. In WG, on the other hand, the fact that words have actors follows completely automatically from the isa hierarchy plus the known properties of actions.

As far as lexicography is concerned, this is a welcome conclusion, because lexicographers have a great deal to say about the actors (i.e. speakers) of words. For instance, a (British) dictionary could reasonably be expected to include the word sidewalk, but to tell us that its speaker is (typically) an American. The WG proposition stating this fact is [23].

[23] actor of sidewalk is an American

Similar treatment can be given to other kinds of social constraints (though I have to admit that I do not at present have any clear idea about how to handle style-levels, such as the difference between try and attempt).

We have already seen how the difference between encyclopedic and 'semantic' information disappears in WG. For example, the meaning of mouse can be given simply as 'a mouse', but then the same knowledge-structure may include propositions about the properties of mice (e.g. that some people are afraid of them, that they squeak, that they like cheese, that they are classified as vermin, that they live in holes in the skirting, and so on).

Another type of information which is excluded from main-stream lexicons is etymology. Such information is particularly important in the case of loanwords, in the sense of words that their users know are borrowed from other languages. Loan-words are often structurally different from other words (e.g. they have features of the pronunciation of the source language which conflict with the normal patterns of the receiving one), and at least some speakers are certainly aware of them as loans. Moreover these speakers know which language it is that they were borrowed from, so the natural assumption in cases like this is that the person has a 'language' proposition about each such word, as in the following:

[24] language of concierge is French

Furthermore I assume that a bilingual speaker may have every word tagged in this way for one language or the other, because there is otherwise no
explanation for the fact that such speakers can either keep their languages separate in speech or mix them up together.

3. Conclusion

In conclusion, linguists and lexicographers seem to have a great deal to learn from one another. Most of the theories that linguists favour cover only a small portion of the information that lexicographers deal with, so it is hard to take seriously their claim to be general theories of ‘knowledge of language’. This rather extraordinary claim relies for its plausibility on a very narrow definition of ‘knowledge of language’ for which there is surprisingly little justification. When the day arrives that linguists recognise that all these other kinds of knowledge ought to be included in their theories of language structure, they will surely turn to lexicographers for more information. However even before then, lexicographers have a lot to offer to linguists, both theoretical and descriptive, because of their vast body of knowledge about strictly intra-linguistic patterns. This can and should be used by linguists both as a data-base against which they can test their theories, and as a yard-stick against which they can measure their progress. Browsing through any modern dictionary is a humbling experience for any linguist who thinks our theories are already quite sophisticated.

As far as lexicographers are concerned, they too have something to learn from improved contacts with linguists. Any experienced grammarian could probably suggest improvements in the grammatical categorisation schemes used by lexicographers, even when these are more sophisticated than the very crude traditional terms like ‘adj’ and ‘trans’. Similar contributions can be expected from the branches of linguistics other than grammar – phoneticians and phonologists can help with pronunciation, sociolinguists and dialectologists can help with regional and other social differences, and so on. The very fact that academic linguists, unlike lexicographers, are not generally working against the clock to meet a deadline enables them to stand back from some of the analytical decisions and see them in a broader perspective which may in the end make the decisions easier to take.

However I think one of the most important contributions that theoretical academic linguists ought to be able to make to the work of the lexicographer is a good general theory of the intellectual structures within which lexicographers work. Notions such as ‘lexical item’, ‘lexical entry’, ‘lexicon’ and even ‘language’ are far more complex and problematic than is generally recognised, and at present I think it would be fair to describe the state of linguistics as a haze of semi-understanding, rather than the crystal clarity that we should all be aspiring to. Theoretical linguists are gradually working towards a shared understanding of these things, and no doubt in the fullness of time a sensible consensus will emerge – in McCawley’s phrase, linguists will have got their act together. Meanwhile lexicographers have to go on making their practical decisions, and for theoretical underpinnings they have the difficult task of
picking and choosing among the available theories of language structure. I hope this paper has succeeded in suggesting that it may be worth looking beyond the easily accessible and much publicised mainstream, to find out what is going on in some minority theories like Word Grammar.

References


Appendix

A check-list of types of lexical fact

The following list aims to cover all the information which could be considered for inclusion in an all-inclusive lexicon. As noted in part 1, mainstream linguists tend to exclude some of these types of information on principle, so their inclusion in this list does not mean that all linguists would agree to including them in their lexicons (or, for that matter, anywhere at all in their lexico-grammars). However, they all constitute facts about words, so the onus is on the excluders to justify their position.

All the types of information included in the list are in fact known about some words by at least some people. (More accurately, they are accepted as true by at least some people – whether or not they are objectively true as statements about the speech of typical people in the community concerned.) Any attempt to model psychological reality must therefore take account of this broad range of knowledge-types; but so must any lexicographer whose purpose is to make accessible all the kinds of knowledge which a typical native speaker has (and which is needed if typical native speech is to be simulated either by a non-native or by a machine).

I have excluded from this list just one type of information which some
researchers believe should be included, namely processing operations specific to particular words. My own view is that such information is not needed, but I recognise that others disagree (e.g. Small, 1980). I leave it to them to make their case. If this approach - 'word expert parsing' - is right, the task of the lexicographer is vastly enlarged and we shall all have to rethink everything. For similar reasons I ignore the opposition between speaking/writing and listening/reading, in the hope that the information in the lexicon is in fact neutral to it - again I recognise that this is somewhat controversial (e.g. Straight, 1986).

The structure of the list is not meant to have much significance - I have argued elsewhere that some of the divisions which I exploit here, such as that between syntax and morphology, are inherently vague and should not be made to carry much weight.

1. Phonology
   - underlying segment structure; or several such structures if allomorphs are stored rather than computed
   - prosodic patterns of word (to the extent that there are no rules for computing these) - i.e. mainly word-stress or tone

2. Morphology
   - structure in terms of morphemes or alternating phonological structures (e.g. Semitic languages)
   - irregular morphological structures linked to particular morpho-syntactic features (i.e. irregular inflections)
   - partial similarities to other words (in the case of derived words or compounds)
   - cliticising properties (i.e. whether or not the word concerned may be used as a clitic or as host of a clitic)

3. Syntax
   - general word-class (e.g. 'verb')
   - sub-class (e.g. 'auxiliary')
   - obligatory morpho-syntactic features (e.g. beware)
   - valency:
     - deviant position of dependent (e.g. someone etc)
     - deviant position of head (e.g. enough)
     - class of dependent (e.g. object of discuss is a noun)
     - class of head (e.g. head of very is an ad-word)
     - morpho-syntactic features of dependent (e.g. objective of folgen is dative)
     - morpho-syntactic features of head (?)
     - lexical identity of dependent (e.g. high-degree-modifier of drunk is blind)
     - lexical identity of head (e.g. head of past-participle sein is sein; neutral prepositional head of foot is on)
     - semantic identity of dependent (e.g. dependent of herd refers to a set of cows)
     - semantic identity of head (e.g. head of each refers to a distributed event)
– semantic identity of dependent if optional and absent (e.g. *He shaved* =
  ‘He shaved himself’)

4. **Semantics**
- name of entity referred to ( = X)
- identity of X (e.g. referent of *me* is speaker)
- more general entities of which X is an instance
- semantic valency of X (i.e. entities to which it is related other than as an
  instance)
- entities which are inherent to X (e.g. purpose of ‘assassinate’)
- entities which are given only as defaults (e.g. ‘theme’ of ‘drink’ is alcohol)
- entities which must be defined by anaphora (e.g. ‘excluded’ of ‘other’,
  ‘known’ of intransitive ‘know’)
- entities which are defined by syntactic dependents (e.g. goal of *reach* is defined
  by direct object)

5. **Context**
- restrictions relating to immediate social structure (e.g. power/solidarity
  markers)
- restrictions relating to style (e.g. ‘formal’, ‘slang’)
- restrictions relating to larger social structure (e.g. speaker classification)
- restrictions relating to discourse structure (e.g. topic-change markers)

6. **Spelling**
- normal orthography
- standard abbreviations or ideographs
- inflectional irregularites in spelling

7. **Etymology and language**
- the language to which the word belongs (in a bilingual dictionary)
- the language from which it is ‘borrowed’
- the word on which it is ‘based’
- the date when it was ‘borrowed’

8. **Usage**
- frequency and familiarity
- age of acquisition
- particular occasions on which the word was used
- clichés containing the word
- taboos

This paper is based on one which was prepared for a workshop ‘On automating the
lexicon’ held on 19–23 May 1986, Marina di Grosseto, Italy. I should like to thank the
organisers of that workshop for asking me to write it, and the other participants,
especially Ron Kaplan, for their helpful comments during the conference. I also want to
thank John Sinclair, Bob Ilson and an anonymous reader for their comments on the
original paper.