DESIGN AND IMPLEMENTATION OF A LEXICAL DATA BASE

Eric Wehrli
Department of Linguistics
U.C.L.A.
405 Hilgard Ave, Los Angeles, CA 90024

ABSTRACT

This paper is concerned with the specifications and the implementation of a particular concept of word-based lexicon to be used for large natural language processing systems such as machine translation systems, and compares it with the morpheme-based conception of the lexicon traditionally assumed in computational linguistics.

It will be argued that, although less concise, a relational word-based lexicon is superior to a morpheme-based lexicon from a theoretical, computational and also practical viewpoint.

INTRODUCTION

It has been traditionally assumed by computational linguists and particularly by designers of large natural language processing systems such as machine translation systems that the lexicon should be limited to lexical information that cannot be derived by rules. According to this view, a lexicon consists of a list of basic morphemes along with irregular or unpredictable words.

In this paper, I would like to reexamine this traditional view of the lexicon and point out some of the problems it faces which seriously question the general adequacy of this model for natural language processing.

As a trade-off between the often conflicting linguistic, computational and also practical considerations, an alternative conception of the lexicon will be discussed, largely based on Jackendoff's (1975) proposal. According to this view, lexical entries are fully-specified but related to one another. First developed for a French parser (cf. Wehrli, 1984), this model has been adopted for an English parser in development, as well as for the prototype of a French-English translation system.

This paper is organized as follows: the first section addresses the general issue of what constitutes a lexical entry as well as the question of the relation between lexicon and morphology from the point of view of both theoretical linguistics and computational linguistics. Section 2 discusses the relational word-based model of the lexicon and the role morphology is assigned in this model. Finally, it spells out some of the details of the implementation of this model.

OVERVIEW OF THE PROBLEM

One of the well-known characteristic features of natural languages is the size and the complexity of their lexicons. This is in sharp constrast with artificial languages, which typically have small lexicons, in most cases made up of simple, unambiguous lexical items. Not only do natural languages have a huge number of lexical elements -- no matter what precise definition of this latter term one chooses -- but these lexical elements can furthermore (i) be ambiguous in several ways (ii) have a non-trivial internal structure, or (iii) be part of compounds or idiomatic expressions, as illustrated in (1)-(4):

(1) ambiguous words:
can, fly, bank, pen, race, etc.
(2) internal structure:
use-ful-ness, mis-understand-ing, lake-s, tri-ed
(3) compounds:
milkman, moonlight, etc.
(4) idiomatic expressions:
to kick the bucket, by and large, to pull someone's leg, etc.

In fact, the notion of word, itself, is not all that clear, as numerous linguists -- theoreticians and/or computational linguists -- have acknowledged. Thus, to take an example from the computational linguistics literature, Kay (1977) notes:

> "In common usage, the term <u>word</u> refers sometimes to sequences of letters that can be bounded by spaces or punctuation marks in a text. According to this view, <u>run</u>, <u>runs</u>, <u>running</u> and <u>ran</u> are different words. But common usage also allows these to count as instances of the same word because they belong to the

same paradigm in English accidence and are listed in the same entry in the dictionary."

Some of these problems, as well as the general question of what constitutes a lexical entry, whether or not lexical items should be related to one another, etc. have been much debated over the last 10 or 15 years within the framework of generative grammar. Considered as a relatively minor appendix of the phrase-structure rule component in the early days of generative grammar, the lexicon became little by little an autonomous component of the grammar with its own specific formalism — lexical entries as matrices of features, as advocated by Chomsky (1965). Finally, it also acquired specific types of rules, the so-called word formation rules (cf. Halle, 1973; Aronoff, 1976; Lieber, 1980; Selkirk, 1983, and others), and lexical redundancy rules (cf. Jackendoff, 1975; Bresnan, 1977).

By and large, there seems to be widespread agreement among linguists that the lexicon should be viewed as the repository of all the idiosyncratic properties of the lexical items of a language (phonological, morphological, syntactic, semantic, etc.). This agreement quickly disappears, however, when it comes to defining what constitutes a lexical item, or, to put it slightly differently, what the lexicon is a list of, and how should it be organized.

Among the many proposals discussed in the linguistic literature, I will consider two radically opposed views that I shall call the morpheme-based and the word-based conceptions of the lexicon[1].

The morpheme-based lexicon corresponds to the traditional derivational view of the lexicon, shared by the structuralist school, many of the generative linguists and virtually all the computational linguists. According to this option, only non-derived morphemes are actually listed in the lexicon, complex words being derived by means of morphological rules. In contrast, in a word-based lexicon a la Jackendoff, all the words (simple and complex) are listed as independent lexical entries, derivational as well as inflectional relations being expressed by means of redundancy rules[2,3].

The crucial distinction between these two views of the lexicon has to do with the role of morphology. The morpheme-based conception of the lexicon advocates a dynamic view of morphology, i.e. a conception according to which "words are generated each time anew" (Hoekstra et al. 1980). This view contrasts with the static conception of morphology assumed in Jackendoff's word-based theory of the lexicon.

Interestingly enough, with the exception of some (usually very small) systems with no morphology at all, all the lexicons in computational linguistic projects seem to assume a dynamic conception of morphology.

The no-morphology option, which can be viewed as an extreme version of the word-based lexicon mentioned above modulo the redundancy rules, has been adopted mostly for convenience by researchers working on parsers for languages fairly uninteresting from the point of view of morphology, e.g. English. It has the non-trivial merit of reducing the lexical analysis to a simple dictionary look-up. Since all flectional forms of a given word are listed independently, all the orthographic words must be present in the lexicon. Thus, this option presents the double advantage of being simple and efficient. The price to pay is fairly high, though, in the sense that the resulting lexicon displays an enormous amount of redundancy: lexical information relevant for a whole class of morphologically related words has to be duplicated for every member of the class. This duplication of information, in turn, makes the task of updating and/or deleting lexical entries much more complex than it should be.

This option is more seriously flawed than just being redundant and space-greedy, though. By ignoring the obvious fact that words in natural languag es do have some internal structure, may belong to declension or conjugation classes, but above all that different orthographical words may in fact realize the same grammatical word in different syntactic environments it fails to be descriptively adequate. Interestingly enough, this inadequacy turns out to have serious consequences. Consider, for example, the case of a translation system. Because a lexicon of this exhaustive list type has no way of representing a notion such as "lexeme", it lacks the proper level for lexical transfer. Thus, if been, was, were, am and be are treated as independant words, what should be their translation, say in French, especially if we assume that the French lexicon is organized on the same model? The point is straightforward: there is no way one can give translation equivalents for orthographic words. Lexical transfer can only be made at the more abstract level of lexeme. The choice of a particular orthographic word to realize this lexeme is strictly language dependent. In the previous example, assuming that, say, were is to be translated as a form of the verbe etre, the choice of the correct flectional form will be governed by various factors and properties of the French sentence. In other words, a transfer lexicon must state the fact that the verb to be is translated in French by etre, rather than the lower level fact that under some circumstances were is translated by etaient.

The problems caused by the size and the complexity of natural language lexicons, as well as the basic inadequacy of the "no morphology" option just described, have been long acknowledged by computational linguists, in particular by those involved in the development of large-scale application programs such as machine translation. It is thus hardly surprising that some version of the morpheme-based lexicon has been the option common to all large natural language systems. There is no doubt that restricting the lexicon to

basic morphemes and deriving all complex words as well as all the inflected forms by morphological rules, reduces substantially the size of the lexicon. This was indeed a crucial issue not so long ago, when computer memory was scarce and expensive.

There are, however, numerous problems — linguistic, computational as well as practical — with the morpheme-based conception of the lexicon. Its inadequacy from a theoretical linguistic point of view has been discussed abundantly in the "lexicalist" literature. See in particular Chomsky (1970), Halle (1973) and Jackendoff (1975). Some of the linguistic problems are summarized below, along with some mentions of computational as well as practical problems inherent to this approach.

First of all, from a conceptual point of view, the adoption of a derivational model of morphology suggests that the derivation of a word is very similar, as a process, to the derivation of a sentence. Such a view, however, fails to recognize some fundamental distinctions between the syntax of words and the syntax of sentences, for instance regarding creativity. Whereas the vast majority of the words we use are fixed expressions that we have heard before, exactly the opposite is true of sentences: most sentences we hear are likely to be novel to us.

Also, given a morpheme-based lexicon, the morphological analysis creates readings of words that do not exist, such as strawberry understood as a compund of the morphemes straw and berry. This is far from being an isolate case, examples like the following are not hard to find:

(5)a. comput-er
b. trans-mission
c. under-stand
d. re-ply
e. hard-ly

The problem with these words is that they are morphologically composed of two or more morphemes, but their meaning is not derivable from the meaning of these morphemes. Notice that listing these words as such in the lexicon is not sufficient. The morphological analysis will still apply, creating an additional reading on the basis of the meaning of its parts. To block this process requires an ad hoc feature, i.e. a specific feature saying that this word should not be analysed any further.

Generally speaking, the morpheme-based lexicon along with its word formation rules, i.e. the rules that govern the combination of morphemes is bound to generate far more words (or readings of words) than what really exists in a particular language. It is clearly the case that only a strict subset of the possible combination of morphemes are actually realized. To put it differently, it confuses the notion of potential word[4] for a language with the notion of actual word[4].

This point was already noticed in Halle (1973), who suggested that in addition to the list of morphemes and the word formation rules which characterize the set of possible words, there must exist a list of actual words which functions as a filter on the output of word formation rules. This filter, in other words, accounts for the difference between potential words and actual words.

The idiosyncratic behaviour of lexical items has been further stressed in "Remarks on Nominalization" where Chomsky convincingly argues that the meaning of derived nominals, such as those in (6), cannot be derived by rules fr om the meaning of its constitutive morphemes. Given the fact that derivational morphology is semantically irregular it should not be handled in the syntax. Chomsky concludes that derived nominals must be listed as such in the lexicon, the relation between verb and nominals beeing captured by lexical redundancy rules.

(6)a. revolve revolution
b. marry marriage
c. do deed
d. act action

It should be noticed that the somewhat erratic and unpredictable morphological relations are not restricted to the domain of what is traditionally called derivation. As Halle points out (p. 6), the whole range of exceptional behaviour observed with derivation can be found with inflection. Halle gives examples of accidental gaps such as defective paradigms, phonological irregularity (accentuation of Russian nouns) and idiosyncratic meaning.

From a computational point of view,' a morpheme-based lexicon has few merits beyond the fact that it is comparatively small in size. In the generation process as well as in the analysis process the lack of clear distinction between possible and actual words makes it unreliable — i.e. one can never be sure that its output is correct. Also, since a large number of morphological rules must systematically be applied to every single word to make sure that all possible readings of each word is taken into consideration, lexical analysis based on such conceptions of the lexicon are bound to be fairly inefficient. Over the years, increasingly sophisticated morphological parsers have been designed, the best examples being Kay's (1977), Karttunen (1983) and Koskeniemmi (1983a,b), but not surprisingly, the efficiency of such systems remain well below the simple dictionary lookup[5].

Also, this model has the dubious property that the retrieval of an irregular form necessitates less computation than the retrieval of a regular form. This is so because unlike regular forms that have to be created/analyzed each time they are used, irregular forms are listed as such in the lexicon. Hence, they can simply be looked up.

This rapid and necessarily incomplete overview of the organization of the lexicon and the role of morphology in theoretical and computational linguistics has emphasized two basic types of requirements: the linguistic requirements which have to do with descriptive adequacy of the model, and the computational requirements which has to do with the efficiency of the process of lexical analysis or generation. In particular, we argued that a lexicon consisting of the list of all the inflected forms without any morphology fails to meet the first requirement, i.e. linguistic adequacy. It was also pointed out that such a model lacks the abstract lexical level which is relevant, for instance, for lexical transfer in translation systems. Although clearly superior to what we called the "no morphology" system, the traditional morpheme-based model runs into numerous problems with respect to both linguistic and computational requirements.

A third type of considerations which are often overlooked in academical discussions, but turns out to be of primary importance for any "real life" system involving a large lexical data base is what I would call "practical requirements" and has to do with the complexity of the task of creating a lexical entry. It can roughly be viewed as a measure of the time it takes to create a new lexical entry, and of the amount of linguistic knowledge that is required to achieve this task.

The relevance of these practical requirements becomes more and more evident as large natural language processing systems are being developed. For instance, a translation system — or any other type of natural language processing program that must be able to handle very large amounts of text — necessitates dictionaries of substantial size, of the order of at least tens of thousands of entries, perhaps even more than 100,000 lexical entries. Needless to say the task of creating as well as the one of updating such huge databases represents an astronomical investment in terms of human resources which cannot be overestimated. Whether it takes an average of, say 3 minutes, to enter a new lexical entry or 30 minutes may not be all that important as long as we are considering lexicons of a few hundred words. It may be the difference between feasible and not feasible when it comes to very big databases[6].

Another important practical issue is the level of linguistic knowledge that is required from the user. Systems which require little technical knowledge are to be preferred to those requiring an extensive amount of linguistic background, everything else being equal. It should be clear, in this respect, that morpheme-based lexicons tend to require more linguistic knowledge from the user than a word-based lexicon, since the user has to specify (i) what the morphological structure of the word is (ii) to what extent the meaning of the word is or is not derived from the meaning of its parts, (iii) what morphophonological rules apply in the derivation of this word.

A RELATIONAL WORD-BASED LEXICON

The traditional view in computational linguistics is to assume some version of the morpheme-based lexicon, coupled with a morphological analyzer/generator. Thus it is assumed that a dynamic morphological process takes place both in the analysis and in the generation of words (i.e. orthographical words). Each time a word is read or heard, it is decomposed into its atomic constituents and each time it is produced it has to be re-created from its atomic constituents[7].

As I pointed out earlier, I don't see any compelling evidence supporting this view other than the simplicity argument. Crucial for this argument, then, is the assumption that the complexity measure is just a measure of the length of the lexicon, i.e. the sum of the symbols contained in the lexicon.

One cannot exclude, though, more sophisticated ways to mesure the complexity of the lexicon. Jackendoff (1975:640) suggests an alternative complexity measure based on "independent information content". Intuitively, the idea is that redundant information that is predictable by the existence of a redundancy rule does not count as independent[8].

Assumimg a strict lexicalist framework a la Jackendoff, we developed a word-based lexical database dubbed relational word-based lexicon (RWL). Essentially, the RWL model is a list-type lexicon with cross references. All the words of the language are listed in such a lexicon and have independent lexical entries. The morphological relations between two or more lexical entries are captured by a complex network of relations. The basic idea underlying this organization is to factor out properties shared by several lexical entries.

To take a simple example, all the morphological forms of the English verb run have a lexical entry. Hence, run, runs, ran and running are listed independently in the lexicon. At the same time, however, these four lexical entries are to be related in some way to express the fact that they are morphologically related, i.e. they belong to the same paradigm. In turns, this has the further advantage of providing a clear definition of the "lexeme", the abstract lexical unit which is relevant, for instance, for lexical transfer, as will be pointed out below.

In contrast with the common use in computational linguistics, in this model morphology is essentially static[9]. By interpreting morphology as relations within the lexical database rather than as a process, we shift some complexity from the parsing algorithm to the lexical data structures. Whether or not this shift is justified from a linguistic point of view is an open question, and I have nothing to say about it here. From a computational point of view, though, this shift has rather interesting consequences.

First of all, it drastically simplifies the task of lexical analysis (or generation), making it a deterministic process — as opposed to a necessarily non-deterministic morphological parser. In fact, it makes lexical analysis rather trivial, equating it with a fairly simple database query. It follows that the process of retrieving an irregular word is identical to the process of retrieving a regular word. The distinction between regular morphological forms and exceptional ones has no effect on the lexical analysis, i.e. on processing. Rather, it affects the complexity measure of the lexicon.

Also, in sharp contrast to what happens with a derivational conception of morphology, in our model, the morphological complexity of a language has very little effect on the efficiency of lexical analysis, which seems essentially correct: speakers of morphologically complex languages do not seem to require significantly more time to parse individual words than speakers of, say, English.

A partial implementation of this relational word-based model of the lexicon has been realized for the parser for French described in Wehrli (1984). This section describes some of the features of this implementation. Only inflection has been implemented, so far. Some aspects of derivational morphology should be added in the near future.

In this implementation, lexical entries are composed of three distinct kinds of objects referred to as words, morpho-syntactic elements and lexemes, cf. figure 1. A word is simply a string of characters, or what is sometimes called an orthographic word. It is linked to a set of morpho-syntactic elements, each one of them specifying a particular grammatical reading of the word. A morpho-syntactic element is a just a particular set of grammatical features such as category, gender, number, person, case, etc. A lexeme contains all the information shared by all the flectional forms of a given lexical item. The lexeme is defined as a set of syntactic and semantic features shared by one or several morpho-syntactic elements. Roughly speaking, it contains the kind of information one expect to find in a standard dictionary entry.

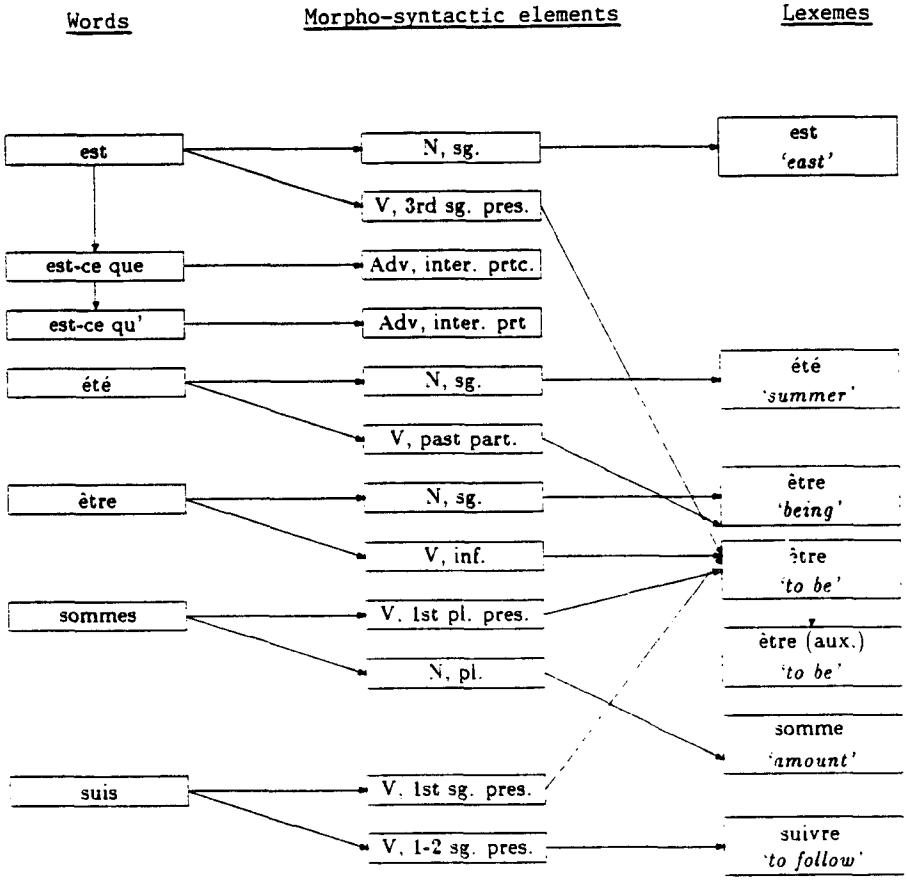

Figure 1: Structure of the lexicon

In relational terms, fully-specified lexical entries are broken into three different relations. The full set of information belonging to a lexical entry can be obtained by intersecting the three relations.

The following example illustrates the structure of the lexical data base and the respective roles of words, morpho-syntactic elements and lexemes. In French, suis is ambiguous. It is the first person singular present tense of the verb etre ('to be'), which, as in English, is both a verb and an auxiliary. But suis is also the first and second person singular present tense of the verb suivre ('to follow'). This information is represented as follows: the lexicon has a word (in the technical sense, i.e. a string of characters) suis associated with two morpho-syntactic elements. The first morpho-syntactic element which bears the features [+V, 1st, sg, present] is linked to a list of two lexemes. One of them contains all the general properties of the verb etre, the other one the information corresponding to the auxiliary reading of etre. As for the second morpho-syntactic element, it bears the features [+V, 1st-2nd, sg, present] and it is related to the lexeme containing the syntactic and semantic features characterizing the verb suivre.

Such an organization allows for a substantial reduction of redundancy. All the different morphological forms of etre, i.e. over 25 different words are ultimately linked to 2 lexemes (verbal and auxiliary readings). Thus, information about subcategorization, selectional restrictions, etc. is specified only once rather than 25 times or more. Naturally, this concentration of the information also simplifies the updating procedure. Also, as we pointed out above, this structure provides a clear definition of "lexeme", the abstract lexical representation, which is the level of representation relevant for transfer in translation systems.

Figure 1, above, illustrates the structure of the lexical database. Boxes stand for the different items (words, morphosyntactic elements, lexemes) and arrows represent the relations between these items. Notice that not all morphosyntactic elements are associated with some lexemes. In fact, there is a lexeme level only for those categories which display morphological variation, i.e. nouns, adjectives, verbs and determiners.

The arrow between the words est and est-ce que expresses the fact that the string est occurs at the initial of the compound est-ce que. This is the way compounds are dealt with in this lexicon. The compound clair de lune ('moonlight') is listed as an independent word -- along with its associated morphosyntactic elements and lexemes -- related to the word clair. The function of this relation is to signal to the analyzer that the word clair is also the first segment of a compound.

Consider the vertical arrow between the lexeme corresponding to the verbal reading of etre ('to be') and the lexeme corresponding to the auxiliary reading of etre. It expresses the fact that a given morphosyntactic element may have several distinct readings (in this case the verbal reading and the auxiliary reading). Thus, morphosyntactic elements can be related not just to one lexeme, but to a list of lexemes.

The role of morphology in Jackendoff's system is double. First, the redundancy rules have a static role, which is to describe morphological patterns in the language, and thus to account for word-structure. In addition to this primary role, morphology also assumes a secondary role, in the sense that it can be used to produce new words or to analyze words that are not present in the lexicon. In this respect, Jackendoff (1975:668) notes, "lexical redundacy rules are learned form generalizations observed in already known lexical items. Once learned, they make it easier to learn new lexical items". In other words, redundancy rules can also function as word formation rules and, hence, have a dynamic function[10].

In our implementation of the relational word-based lexicon, morphology has also a double function. On the one hand, morphological relations are embedded in the structure of the database itself and, roughly, correspond to Jackendoff's redundancy rules in their static role. On the other hand, morphological rules are considered as "learning rules", i.e. as devices which facilitate the acquisition of the paradigm of the inflected forms of a new lexeme. As such, morphological rules apply when a new word is entered in the lexicon. Their role is to help and assist the user in his/her task of entering new lexical entries. For example, if the infinitival form of a verb is entered, the morphological rules are used to create all the inflected forms, in an interactive session. So, for instance, the system first considers the verb to be morphologically regular. If so, that is if the user confirms this hypothesis, the system generates all the inflected forms without further assistance. If the answer is no, the system will try another hypothesis, looking for subregularities.

Our relational word-based lexicon was first implemented on a relational database system on a VAX-780. However, for efficiency reasons, it was transfered to a more conventional system using indexed sequential and direct access files. In its present implementation, on a VAX-750, words and morphosyntactic elements are stored in indexed sequential files, lexemes in direct access files. In other words, the lexicon is entirely stored in external files, which can be expanded, practically without affecting the efficiency of the system. A set of menu-oriented procedures allow the user to interact with the lexical data base, to either insert, delete, update or just visualize words and their lexical specifications.

CONCLUSION

Several important issues have been discussed in this paper, regarding the structure and the function of the lexicon, as well as the role of morphology. We first pointed out the important role of morphology and showed that it cannot be dispensed with, even in processing systems with no particular psychological claim. Hence, an exhaustive list of all the orthographic forms of English words cannot stand for an adequate lexicon of English.

Turning then to what appears to be the traditional conception of morphology in computational linguistics, we showed that a morpheme-based lexicon, along with a derivational morphological component faces a variety of serious problems, including its inability to distinguish actual words from potential words, its inability to express partial morphological or semantic relations, as well as its inherent inefficiency and often lack of reliability.

The success of this traditional conception of the lexicon in computational linguistics must probably be attributed to its relative conciseness. However, alternative ways to evaluate the complexity of lexical entries, i.e. Jackendoff's independent information content, as well as the emergence of cheap and abundant memory have drastically modify this state of affair, and open new perspectives more in line with current research in theoretical linguistics.

To the traditional view, we opposed a relational word-based lexicon, along the lines of Jackendoff's (1975) proposal, where morphology can be viewed, in part, as relations among lexical entries. Simple words, complex words, compounds, etc., are all listed in our lexicon. But lexical entries which belong to a same paradigm are related to the same lexeme. Rather than deriving or analyzing words each time they are used, morphological rules only serve when a new word occurs.

FOOTNOTES

1. One might think of compromises between these two options, such as, for instance, the stem-based lexicon argued for in Anderson (1982), where lexical entries consists of stems rather than morphemes, and an independent morphological component is responsible for the derivation of inflectional forms.
   Aronoff's (1976) proposal can also be viewed as a compromise solution. See footnote 2.

2. It should be pointed out that other word-based theories have been proposed. For instance, Aronoff (1976) argues for a word-based lexicon where only words which are atomic or exceptional in one way or another are entered in the lexicon.

3. In this paper, I will simply consider inflectional morphology as the adunction to words of affixes which only modify features such as tense, person, number, gender, case, etc. as in read-s, read-ing, book-s. Derivational morphology, on the other hand, deals with the addition of affixes which can modify the meaning of the word, and very often its categorial status, e.g. use-ful, use-ful-ness, hard-ly.

4. Potential words are words that are well-formed with respect to word formation rules, whereas the actual words are the those potential words that are realized in this language. To give an example, both arrival and arrivation are potential English words, but only the second happens to be an actual English word.

5. For instance, Koskeniemmi (1983b) mentions an average of 100 milliseconds per words on a DEC-20.

6. This figure is indeed very conservative. Slocum (1982:8) reports that the cost of writing a dictionary entry for the TAUM-Aviation project was estimated at 3.75 man-hours...

7. This conception is yet another example of the "historicist approach" typical of classical transformational generative grammar, which assumes that synchronic processes recapitulates many of the diachronic developments.

8. The following is an approximation of how independent information can be measu red:
   "(Information measure)
   Given a fully specified leixcal entry W to be introduced into the lexicon, the independent information it adds to the lexicon is
   (a) the information that W exists in the lexicon, i.e. that W is a word of the language; plus
   (b) all the information in W which cannot be predicted by the existence of some redundancy rule R which permits W to be partially described in terms of information already in the lexicon; plus
   (c) the cost of referring to the redundancy rule R.

9. It will be argued below that morphology has a secondary role, which is to facilitate the acquisition of new words.

10. In the conclusion of his "Prolegomena" Halle also mentions the possibility th at word formation rules be used when the speaker hears an unfamiliar word or when he uses a word freely invented.

11. From a psychological point of view, it could also be argued that morphology facilitates memorization.

REFERENCES

Anderson, S. R. (1982). "Where is morphology?", Linguistic Inquiry.

Aronoff, M. (1976). Word Fromation in Generative Grammar, Linguistic Inquiry Monograph One, MIT Press.

Bresnan, J. (1977). "A realistic transformational grammar", in Halle, M., J. Bresnan and G.A. Miller (eds.) Linguistic Theory and Psychological Reality, MIT Press.

Chomsky, N. (1957). Syntactic Structures, Mouton.

Chomsky, N. (1965). Aspects of the Theory of Syntax, MIT Press.

Chomsky, N. (1970). "Remarks on nominalization", Studies on Semantics in Generative Grammar, Mouton.

Halle, M. (1973). "Prolegomena to a theory of word formation", Linguistic Inquiry, 4.1. pp. 3-16.

Hoekstra, T., H. van der Hulst and M. Moortgat (1983). Lexical Grammar, Foris.

Jackendoff, R. (1975). "Morphological and semantic regularities in the lexicon", Language 51.3, pp. 639-671.

Karttunen, L. (1983). "KIMMO: A general morphological processor". Texas Linguistic Forum, No. 22, pp. 165-228.

Kay, M. (1977). "Morphological and syntactic analysis", in A. Zampoli (ed.) Linguistic Structures Processing, North-Holland.

Koskenniemi, K. (1983a). Two-Level Morphology: A General Computational Model For Word-Form Recognition And Production, Publications No 11, University of Helsinki.

Koskenniemi, K. (1983b). "Two-Level Model for Morphological Analysis", Proceedings of the Eighth International Joint Conference on Artificial Intelligence, pp. 683-685, William Kaufmann, Inc.

Lieber, R. (1980). On the Organization of the Lexicon, Ph.D. Dissertation, MIT.

Selkirk, E. (1982). The Syntax of Words. Linguistic Inquiry Monograph Seven, MIT Press.

Slocum, J. (1981). "Machine translation: its history, current status and future prospects", mimeo, University of Texas.

Wehrli, E. (1984). "A Government-Binding parser for French", working paper no 48, ISSCO-Geneva University.