Linguistic Inquiry
Monograph Fourteen

On the Definition of Word

Anna Maria Di Sciullo and
Edwin Williams

On the Definition of Word

Anna-Maria Di Sciullo
and Edwin Williams

The MIT Press
Cambridge, Massachusetts
London, England

This book was set in Times New Roman by Asco Trade Typesetting Ltd., Hong Kong, and printed and bound by Halliday Lithograph in the United States of America.

Library of Congress Cataloging-in-Publication Data

Di Sciullo, Anna-Maria, 1951–
On the definition of word.

(Linguistic inquiry monographs; 14)
Bibliography: p.
Includes index.
1. Word (Linguistics) I. Williams, Edwin. II. Title. III. Series.
P128.W67D5 1987 415 86-27206
ISBN 0-262-04091-3
ISBN 0-262-54047-9 (pbk.)

Linguistic Inquiry Monographs
Samuel Jay Keyser, general editor

1. *Word Formation in Generative Grammar*
   Mark Aronoff
2. *$\bar{X}$ Syntax: A Study of Phrase Structure*
   Ray Jackendoff
3. *Recent Transformational Studies in European Languages*
   Samuel Jay Keyser, editor
4. *Studies in Abstract Phonology*
   Edmund Gussmann
5. *An Encyclopedia of AUX: A Study in Cross-Linguistic Equivalence*
   Susan Steele
6. *Some Concepts and Consequences of the Theory of Government and Binding*
   Noam Chomsky
7. *The Syntax of Words*
   Elisabeth Selkirk
8. *Syllable Structure and Stress in Spanish: A Nonlinear Analysis*
   James W. Harris
9. *CV Phonology: A Generative Theory of the Syllable*
   George N. Clements and Samuel Jay Keyser
10. *On the Nature of Grammatical Relations*
    Alec Marantz
11. *A Grammar of Anaphora*
    Joseph Aoun
12. *Logical Form: Its Structure and Derivation*
    Robert May
13. *Barriers*
    Noam Chomsky
14. *On the Definition of Word*
    Anna-Maria Di Sciullo and Edwin Williams

Series Foreword

We are pleased to present this monograph as the fourteenth in the series *Linguistic Inquiry Monographs*. These monographs will present new and original research beyond the scope of the article, and we hope they will benefit our field by bringing to it perspectives that will stimulate further research and insight.

Originally published in limited edition, the *Linguistic Inquiry Monograph* series is now available on a much wider scale. This change is due to the great interest engendered by the series and the needs of a growing readership. The editors wish to thank the readers for their support and welcome suggestions about future directions the series might take.

Samuel Jay Keyser
for the Editorial Board

On the Definition of Word

Contents

The Notion "Word"

There are three different ideas of what a word is. Our purpose here is to sort them out, and in this introductory section we will do this in a preliminary way. The notion central to this text concerns the *form* of a certain set of objects; the definition of this set is the theory of morphological form. The theory defines the set by means of a set of atoms (morphemes) and rules of combination (affixing, compounding). We will call the members of this set *morphological objects* and assert that the central task of morphology is to determine the laws of form that determine membership in this set. Just as morphology has atoms, so does syntax, and words are commonly taken to be the atoms of syntax. We will call words in this sense *syntactic atoms*. This notion of word is conceptually and factually distinct from that of word as "morphological object." We will discuss an important class of items that are syntactic atoms (insertable in X^0 slots in syntactic structures) but that do not have morphological form (in fact they have syntactic form). Finally, the third notion of word we want to discuss is the one from which most confusion about words derives—the notion of words as the "listed" units of language. For this notion of word, which we believe to be of no interest to the grammarian (though perhaps it is to the psychologist), we coin the term *listeme*. We will show that the listemes of a language correspond to neither the morphological objects nor the syntactic atoms of a language.

Morphology and syntax are similar in that each has a set of atoms and some rules of combination. It is our thesis that the difference between these two subtheories is exactly a difference in the atoms and in the properties of the rules of formation. There are of course other imaginable ideas of what constitutes the difference between syntax and morphology. For example, one might conceive of morphology as a theory of the lexicon, an innocuous-sounding conception but one from which we feel flows a great deal of confusion. Another perhaps related idea of the difference is that

syntax is a characterization of the "productive" aspects of language and morphology a characterization of the semi- or unproductive aspects. These views tend to equate *listeme* and *morphological object*. Under such a view, the following diagram characterizes the "ideal" language:

	Word	Phrase
Listed	yes	no
Unlisted	no	yes

Of course there are exceptions in every cell of this diagram. The question is, Are these exceptions deviations from an ideal, or is the ideal misconceived in the first place? A difficult question.

For example, there are vast veins of productive (and so we assume unlisted) morphology, such as the words ending in *-ness* in English. In addition, there are unproductive and therefore listed veins of idiomatic syntactic constructions, such as the verb-particle construction in English. In fact we can show that all cells in this diagram are substantially filled and should be marked "yes."

But our objection to the diagram is more fundamental than these factual observations would suggest. The diagram itself, however the cells are filled, is an artifact of the confusion we seek to address.

The distinction between word and phrase is a distinction in the theory of grammar. The listed/unlisted distinction has nothing to do with grammar. Syntax and morphology are both recursive definitions of sets of objects—but of different sets, with different atoms and different rules of combination. These are the only differences the grammarian need acknowledge.

The most immediate consequence of this view is that productivity and listedness are not grammatical concepts. We will explore this in chapter 1, where we will examine the property of "listedness." A second consequence is that the lexicalist hypothesis (which we call here the *thesis of the atomicity of words*) is not a principle of grammar but rather a consequence of the conception that grammar contains two subparts, with different atoms and different rules of formation. We will take this up in chapters 3 and 4, which concern the notion "syntactic atom." In chapter 2 we will present a substantive account of the laws and rules of word formation, which we take to constitute the notion "morphological object."

In sum, then, we postulate three empirically and conceptually distinct notions of word: listeme, syntactic atom, and morphological object. In fact, though, there is a fourth, which we will essentially ignore here: the notion of phonological word. We take it up briefly at the end of chapter 4 mainly to draw the contrast with the other senses of word.

Chapter 1

Listeme: The Property of Listedness

Knowledge of language involves in some way a knowledge of particular linguistic objects—for example, the word *transmission* and the knowledge that it (1) has a certain morphological form and (2) refers to a part of a car; that *take to task* has a certain syntactic form and means "rebuke". To the extent that an object does not have the form or interpretation specified by the recursive definitions of the objects of the language, that object and its properties must be "memorized." We have dubbed such memorized objects *listemes*, and this property of being memorized, *listedness*. Our overall point is that listedness is no more intrinsically characteristic of words than it is of phrases. Some words and some phrases are listed, but infinitely many of each are not.

If conceived of as the set of listemes, the lexicon, is incredibly boring by its very nature. It contains objects of no single specifiable type (words, VPs, morphemes, perhaps intonation patterns, and so on), and those objects that it does contain are there because they fail to conform to interesting laws. The lexicon is like a prison—it contains only the lawless, and the only thing that its inmates have in common is lawlessness.

This view of the lexicon is quite unfriendly to most current proposals, which by and large can be traced to the influential view of Jackendoff (1975). According to Jackendoff, all words of a language are listed in the lexicon, whether or not they conform completely to the laws of form and meaning for words. The rules of morphology are conceived of as redundancy rules, by means of which the "cost" of a lexical item is computed. Those that are totally predictable will have no cost.

We do not take issue with the view that the "cost" of a word is a function of its deviation from the rules of morphology. As noted, a structured item is easier to memorize than an unstructured one. But we do take issue with the idea that the rules of morphology are essentially redundancy rules over a finite set of objects in a way that syntactic rules are not. To us this makes

no more sense than to say that VP → V NP is a redundancy rule over the set of VPs, most but of course not all of which have zero redundancy.

Jackendoff acknowledges that there are phrases in the lexicon (idioms) and that "possible words" that are not in the lexicon can be created and used, but these aberrations do not deter him from conceiving of the lexicon as a list of all the words of a language, and of the rules of morphology as a "theory" of that list.

A related view that we reject is the idea that "productivity" is characteristic of syntax and "unproductivity" is characteristic of morphology. Fabb (1984, 38), for example, explicitly considers this a criterial difference, and many others implicitly consider it as such. Selkirk (1981), for example, proposes that productive compounds are derived in syntax and unproductive or idiosyncratic ones are listed in the lexicon. In our view, to the extent that productivity is phenomenally perceived to distinguish syntax and morphology, this is something to explain, not something that follows from the intrinsic nature of these two systems, and the explanation is not all that interesting (see section 1.2.3).

Another related view that we reject is the idea that the lexicon has structure. As mentioned, it is simply a collection of the lawless, and there neither can nor should be a theory directly about it, for it can only be understood in terms of the laws it fails to obey. This is not to say that the space of words in a language is not structured—in fact the space of words has a rich structure, imposed first by the rules of word formation and second by the paradigmatic matrices that words enter into (see Williams 1981). But the lexicon contains only some of the words that enter into this structure (the ones that do it least well), and it contains much else besides. In sum we reject the idea that listedness is a grammatical property—the lexicon is a collection of semigrammatical objects, some of them words and others phrases. The set of listed items has no structure, and the property of being a member of this set is no more essential to the nature of words than it is to the nature of phrases.

Of course this is not to say that knowledge of the listed items of a language is not part of knowledge of that language. Rather, if we think of the rules of formation for words and phrases as defining the grammatical items of the language, then we might regard the lexicon as containing a finite list of some semigrammatical objects that are a part of the language.

In the next two sections we will explore somewhat the reasons for distinguishing the lexicon from the space of words of a language. We will concentrate on the two important types of case for which these two

concepts fail to coincide: (1) the listed syntactic objects and (2) the unlisted (and unlistable) morphological objects.

1.1 Listed Syntactic Objects

The listed syntactic objects are the idioms. Although these are like some words in that their meanings cannot be compositionally computed, this very feature makes them unlike most words.

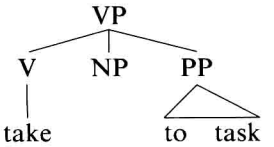
As far as we can see, there is nothing more to say about them than that (1) they are syntactic objects and (2) they are listed because of their failure to have a predictable property (usually their meaning).

It is not trivial to say that an item is a syntactic object: it means that the item is a syntactic unit of some kind—an NP, VP, and so on. It is certainly conceivable that some concatenation of words that was not a unit could be an idiom, or that some particular string of words, such as *park saw* in *The man in the park saw the woman*, could contribute an unpredictable element of meaning to a sentence, but idioms do not work like this—they are always units.

They do not always look like units. For example, the VP *take NP to task* looks like a discontinuous unit, wrapped around the object NP—in fact it has been proposed (by Emonds (1969) for one) that *take to task* is a complex V and not a VP at all (with *to task* extraposed by a rule).

This description, which would enormously complicate the morphological notion "verb," is not forced on us, because we can assign the idiom the following structure:

(1)



This idiom contains an unfilled position, the NP object; nevertheless, it is formally a VP and functions as one when inserted for the node VP in a sentence.

There are many such VPs in English (*take in hand*, *push too far*), all of whose properties follow if they are viewed simply as listed VPs. There are probably as many such VPs as there are noncompositional words ending in *-ion*, so this should not be viewed as a minor appendix to the dictionary.

Another set of such VPs in English consists of verb-particle constructions (*look up*, *throw up*, and so on), which are listed VPs consisting of a V, a

particle, and an (optional) open NP position. We will discuss this system shortly.

The discontinuity these VPs exhibit follows from allowing one free open position. This need not be in the object position—it can be inside the direct object position (for example, *bite NP's head off*). In general it can probably be anywhere.

French exhibits a more interesting discontinuity—idioms that include clitics such as *en* (see Di Sciullo 1983):

(2) [en mettre plein la vue]$_{VP}$
'to impress greatly'
[en voir de toutes les couleurs]$_{VP}$
'to have a hard time'

These clitics can wind up in S-Structure separated from the rest of the idiom, as in the following examples:

(3) Jean en voit de toutes les couleurs
Jean en a vu de toutes les couleurs
Jean en fait voir de toutes les couleurs

The separation is exactly what one would expect if the VPs in (2) were inserted for VP nodes in S and cliticization were then to apply. Incidentally, these apparently provide strong evidence for a cliticization rule—the clitic attaches to the first finite verb, whether or not it is part of the idiom. Again, the discontinuity is no obstacle to calling these syntactic units, specifically, VPs.

Calling idioms listed syntactic units means not only that they will be units in the first place but also that they will have the internal structure of other syntactic units in the language and will behave as other units do in syntax; and, as is clear from these examples, this is the case.

We are further led by this view to expect to find listed syntactic units of all kinds—NP, AP, PP, VP, S—and we do:

(4) AP all wet
PP in the dark about NP
S the cat has got NP's tongue
N′ that son of a bitch
NP The Big Apple

The great wealth of such expressions in languages substantiates half of the view that there is nothing special about listed words. In fact there are listemes among all the syntactic categories, perhaps as many as there are words. Further, these listemes have quite regular internal syntax, syntax of the kind given by phrasal syntax, not by morphology.

1.2 Unlisted Morphological Objects

The second part of our argument that there is no correspondence between listedness and morphological objecthood is the demonstration that there are unlisted morphological objects. Actually this is quite widely recognized; all theories of morphology acknowledge, for example, the ability of speakers to make up new words. Thus Halle (1973) cites the series of words *anti-missile missile*, *anti-anti-missile missile missile*, and so on—an infinite series of words, each with a determinate meaning different from that of all the others.

Most views, though, attach grammatical significance to the use of a new word—when a speaker makes up a new word, he changes his grammar by adding the word to his lexicon, even if the word's form and meaning are completely determined by regular rule (see, for example, Jackendoff 1975). But in our view the listedness of a regular form is of no grammatical significance, and whether or not it is listed will vary from speaker to speaker, determined by such factors as its frequency of use in the speaker's daily life.

Most views of morphology distinguish it from syntax in some way that has to do with productivity: use of a new (but regular) syntactic phrase does not result in that phrase being listed in the dictionary, but the use of a new word will. This, so the story goes, is because syntax is inherently productive, whereas morphology is inherently nonproductive or only marginally or spottily productive. This difference is related to a second difference: "blocking" of one form by another (in the sense of Aronoff 1976) only obtains among morphological objects, not among syntactic objects.

In the next two sections we will argue that neither of these distinctions between syntax and morphology is real—that both systems are productive in the same way and that blocking, to the extent that it holds at all, holds in both systems and in fact holds across both systems.

1.2.1 Productivity

It is often noted (and great consequence is attached to it) that *-ness* is more productive than *-ion*. Thus practically any adjective can have *-ness* attached, but only a few select verbs have nominalizations in *-ion* (**breakion*, **cution*, **bition*, **killion*). These two affixes are said to differ in productivity. We might imagine assigning a productivity index to each affix, where we arrive at the index by dividing the number of forms that the affix actually attaches to by the total number of items belonging to the part of speech (C) that the item attaches to:

$$(5)\ P_{af} = \frac{\text{size (X-af)}}{\text{size (C)}}$$

But as Aronoff (1976) understood, the productivity of an affix is not uniform across an entire part of speech; *-ion*, for example, is highly productive (in fact P = 1) for verbs of the form X-*ate* (*emancipate*, *calibrate*). This provokes us to ask, Why compute productivity over a part of speech? Why not a smaller domain (verbs ending in *-ate*) or a larger domain (the entire lexicon)? Is there any answer to this that is not arbitrary?

Suppose that we compute productivity within the contextual restrictions of the affix itself. Thus *-ness* is restricted to As, so the productivity is computed with respect to size (A). But *-ion* is restricted to the Latinate subvocabulary in English. In that subvocabulary *-ion* is extremely productive; perhaps P = 1.

It may be objected that this computation of P_{-ion} is artificial because the only way to identify the Latinate subvocabulary is to see what *-ion* attaches to in the first place. But in this case the charge is not true; there are other ways to identify the subvocabulary. For example, consider the class of words that *-ive* attaches to. The affix *-ion* is 100 percent productive across that independently identified class. Actually the Latinate vocabulary is a closed class to start with, but *-ion* is extremely productive within that class, just as *-ness* is productive across its class (the class of all As). The only difference between *-ion* and *-ness* is that *-ness* is initially defined for a larger class.

In fact *-ion* is productive in the most basic sense of the word—it can be used to make up new words. Of course the base word must be of the right type, which happens to be a relatively closed class to begin with, so most of the possible *-ion* attachments already "exist." But if one encounters a new verb ending in *-ate* (say, *lucubrate*), one does not have to guess, one knows, that a word can be derived from it by adding *-ion* (*lucubration*).

It may seem that we want to regard *-ion* as productive. But productive compared to what? Compared to the class of words it is defined to attach to. But this class is defined in terms of a nonuniversal rule feature, the feature +Latinate. If productivity can be defined with respect to such features, can any affix be less than 100 percent productive? The answer is not obvious.

We have so far drawn into question any firm conclusions based on differences in productivity among rules of morphology. What about differences in productivity between morphology and syntax? To illuminate this discussion, we will examine two subsystems of English verbs.

One subsystem is the Latinate subsystem, consisting of Latinate prefixes (*de-*, *in-*, *re-*, *sub-*, and so on) and stems (*-duct*, *-ject*, *-fer*, *-sist*); our discussion is based on Aronoff 1976:

	-ject	-sist	-fer	-duct
de-	*	*	*	*
in-	*	*	*	*
re-	*	*	*	–
sub-	*	*	–	*

As the chart reveals, this space of words is quite dense, at least for this (carefully chosen) sample. Does this mean that this morphology is productive? It is difficult to say, for two reasons. First, there are a finite number of Latinate prefixes and stems, so there are only so many possible forms of this kind, and most of them exist. Second, because there is no productive semantics for the class of forms, they must all be listed in any case. A good example is *subduct*—this word was probably invented in the service of the new theory of plate tectonics, and its meaning is technical and was invented simultaneously (similar to *subjacency* in linguistics).

We are not as interested in determining whether this system is productive as we are in comparing it with another system in English, the verb-particle system (*look up*) mentioned in section 1.1. This system is undoubtedly phrasal—a VP, with the head verb on the left: the left element takes inflection (*looks up*); the right element can be modified by adverbs (*look right up*); syntactic elements can intervene (*look it up*); and so on. Consider the following array:

	give	throw	stand	look	call
up	*	*	*	*	*
down	–	·	·	·	*
in	*	·	*	*	*
out	*	*	*	*	*

Consider the similarities between this chart and the one of Latinate verbs. Again the space is dense—every form for these verbs and particles exists but one. Again there is no rule for giving the meaning of most of the forms ("*" designates noncompositional forms and "." compositional forms). And again the particles seem to be restricted to a subvocabulary—a particle with a Latinate verb is uncommon (**donate up*).

There seems to be no essential difference between the Latinate prefix-stem system and the verb-particle system with respect to either productivity or compositionality. But one of these systems is lexical and the other syntactic; that is, one is a part of the definition of English word, and the