I Introduction

This book is concerned with the numeral systems of natural languages – with the ways, that is, in which people in various parts of the world count with words. In this Introduction I shall give the reader an idea of the central purpose of the book and draw an outline of the principal methodological assumptions which linguists bring to the study of language and those which I have brought to the study of the numeral systems of languages. In subsequent chapters I shall describe in some detail the numeral systems of, first, English, and then a variety of other, more or less ‘exotic’ languages. The motivation for this book is to discover the deep and general properties exhibited by numeral systems, rather than just to look at the superficial facts. Learning to count up to a given point in as many languages as possible is a pastime that has given pleasure to many people, but our purpose is to delve deeper, to investigate the special ways in which linguistic counting systems are organized. To get an inkling of what we are up to, try to list as many possible ways in which the sequence of words one, two, three, four, five, six, seven, eight, nine, ten, eleven, twelve, thirteen, fourteen differs from the sequence of words that is pronounced as ay, bee, cee, dee, ee, eff, gee, aitch, eye, jay, kay, ell, emm, enn. Do not be content to list just the most obvious differences; say as much as you can, drawing on all of your knowledge of the use of these words and the use of corresponding words in other languages. If that exercise strikes the reader as sufficiently intriguing or bizarre, he may read the rest of this book.

Consider the list of words one, two, three, four, five, six, seven, eight, nine. It is entirely natural to us to think of this list of words as a series or progression; it is equally obvious that the fact that it is a series or progression has nothing to do with the phonetic or phonological properties of the words. A foreigner learning English could no more predict that the next word is pronounced ten than we can predict what the name will be of the tenth planet in the solar system, if one is ever
2 1 Introduction

discovered. The knowledge that the other nine are called Mercury, Venus, Earth, etc., gives us no help at all. But the foreigner learning English will have very definite expectations about the reference of the next item in the list of English number words given above. He will expect it to refer to a number ‘one greater’ than did the last term given. It is, then, more appropriate to say that the list of English words one, . . . , nine represents a series or progression and that this progression is one of distinct, yet closely related number concepts.

A foreigner learning English is not at the same time learning arithmetic. He already has a clear idea of the relationships between the concepts underlying the English words eight, nine, ten; he is merely learning new names for them. In his own language he uses different names. Every known language has a way of naming at least a few numbers. The notion of numenation and the concepts of particular numbers seem to be universal, i.e. a comprehension of them is accessible to every healthy adult. A theory of language must contain the means for describing how each particular language associates arbitrary phonological sequences (words) with these universal concepts. We need a natural and ‘neutral’ notation, independent of any particular language, for representing number concepts.

The most natural and neutral form of representation to adopt is one from which one may actually ‘count off’ the number which is being represented in any given case. Thus, to represent the concept underlying the English word three, for example, I shall use three marks ///; similarly, the meaning of the English word nine will be represented as ////////. The actual shape of the marks has no significance; all that is important is the number of marks in a given representation. I shall call these sets of marks ‘semantic representations’. The particular form selected for the semantic representations of numbers is entirely appropriate to the nature of the arithmetical operations, such as subtraction, multiplication, etc., in which numbers may be involved. All such operations can be defined in terms of the basic arithmetical operation, namely ‘adding one’, and this basic operation can be represented simply by the analagical operation, performed on the semantic representations we are proposing, of ‘making another mark’. Thus arithmetical operations involving number concepts can be described quite simply if we adopt these semantic representations. These semantic representations are linguistic universals. They will figure in the descriptions of any language which deals with numbers, that is to say of every language.
Just as human beings are universally capable of distinguishing between different number concepts, e.g. between ///// and ///, so I assume that they are also universally capable of distinguishing numbers as a class of concepts from all other classes of objects, abstract or otherwise, in the universe. Number concepts are the only ones that may be combined with each other by arithmetical operations such as addition, subtraction, multiplication, and exponentiation. Although little can be said with clarity on the subject, it is certainly clear that number concepts interact with other concepts in a unique way. Colour concepts, for example, do not interact with the concepts of concrete objects in the same way as do number concepts.

Similarly in most, if not all of the world’s languages, the phonological sequences (words and longer expressions) which are used to represent number concepts form a class with unique distributional properties. In English, for example, the set of contexts where one might possibly use a number expression, e.g. six, twenty nine, eight hundred million and two, etc., is not just that set of contexts in which one might also use an adjective, or a noun, or a verb, or a member of any other recognizable syntactic class. The word which in English comes closest to the number expressions in its distributional properties is probably many, but there are contexts in which many may be used but where a number expression would be unacceptable. Compare too many, very many, and how many with *too six, *very six and *how six. Again, there are contexts where number expressions may occur, but where many may not occur. Compare exactly six, less than six and almost six with *exactly many, *less than many, and *almost many.

We should not be surprised that a unique class of concepts is correlated by a language, and probably by all languages, with a unique class of expressions. This amounts to saying that probably all languages have as a component something that can be called a numeral system, a system distinct from all other systems in the same language. Let us postulate the abstract category NUMBER, representing in the theory of languages generally all that is peculiar to numeral systems as distinct from other language systems, such as the Verb system or the Noun system. Thus to associate a particular phonological form, say English seven, in some way with the universal category NUMBER is to express the fact that this form is an element in a numeral system. Similarly, to associate a particular semantic representation, say /////, with NUMBER is to express the fact that it too is an element in a numeral
system. The purpose of this book is to define the correlation between
the universal class of number concepts and the class of number expres-
sions found in several languages, and, moreover, to define the correlation
in a way that seems likely to be appropriate for all languages. In short,
we are setting out to express significant generalizations about NUMBER.

Now it can be argued that the class of number expressions in any
given language is infinite. Intuitions of language users differ on the
matter of whether the set of number expressions in their language is
infinite. The crux of the matter is the question whether the names for
very high numbers are in fact wellformed. In English, for example, the
expression two billion billion, five hundred and five may be felt by some
speakers to be quite wellformed, though of course unlikely to be
observed, whereas other speakers may object that it is not wellformed.
It is significant, however, that there are some high number expressions
which are unacceptable for all speakers, e.g. *five trillion thousand and
ten. We shall see later that we can formulate statements which predict
accurately that expressions of this latter sort are illformed. These
statements will, furthermore, also correctly characterize as illformed
certain lower number expressions such as *two hundred hundred and
*twenty ten, over which there is no disagreement at all between speakers
of English. In short, the unacceptability of some high number names
seems to be systematic, whereas there does not appear to be such a
systematic way of predicting the acceptability of the unclear cases in
which there is disagreement between native speakers of the language.
It will become obvious as we proceed that the particular systematic
characteristics which are evident in natural language number-name
systems tend to project the existence of infinite sets of number-names
and a higher limit to the value of wellformed number-names can only
be stated in a fairly ad hoc, arbitrary manner. We need, therefore, some
device which is capable of generating, or specifying precisely, the entire
set of number-names in a language without actually listing them. The
idea of simply listing all the number-names in a language is unaccept-
able. To do so would be to miss the obvious generalizations which hold
true about number-name systems, as we shall see; we are committed,
in fact, by the nature of the data to describing an infinite set and it is
obviously impossible to make a list of the members of an infinite set.

The set of number concepts is, also, of course, infinite. Therefore we
definitely cannot proceed by simply listing the semantic representations
of all numbers and correlating each item on the list with its appropriate
phonological form. We must again postulate some device which is able to generate, or define, the class of semantic representations for number concepts without listing all its members. What we are setting out to do might be represented schematically in a variety of ways. Diagrams (1)–(4) show some of these ways. In diagram (1) separate devices are postulated to define the two infinite sets with which we are dealing and the outputs of these devices are related by a set of statements represented by the downward ‘hand’ sign. In diagrams (2), (3), and (4), on the other hand, a single device is postulated. In (2) the device is one which specifies an infinite set of number-names for some language and uses this set as the basis for specification of the universal infinite set of semantic representations. In (3) the device generates the universal and infinite set of semantic representations for numbers, which is used as a basis for the specification of the infinite set of number-names in the language we happen to be dealing with. In (4) the device specifies an infinite set of abstract objects which are neither semantic representations nor the phonological specifications of number-names in any language. This set is represented by the letter X. X is used as a basis for the specification of both the infinite sets of semantic and phonological representations.

It should be obvious that the ‘hand’ signs in diagrams (1)–(4) do not necessarily stand for the same thing. They symbolize only whatever sets of statements are necessary to relate the sets and/or devices between
which they are placed in a given diagram. Consider, for example, the
set of statements needed to specify the set of number-expressions in a
particular language using as a basis the already defined set of universal
semantic representations (i.e. the set of statements represented by the
downward ‘hand’ in diagram (3)). This set of statements is not likely to
look the same as the set of statements needed to specify the universal set
of semantic representations using as a basis the already specified set of
number-expressions from some language (i.e. the set of statements
represented by the downward ‘hand’ in diagram (2)). Some of the
‘hands’ in diagrams (1)–(4) do represent identical sets of statements,
e.g. the leftmost hand in (3) and the top leftmost hand in (1).

The descriptive frameworks schematically represented in diagrams
(1)–(4) are not the only possible ones, but they are among the most
obvious. Two of them, namely (3) and (4), correspond to possible views
of what a linguistic description should look like, the rival merits of
which are being hotly debated at the time of writing. Diagrams (1)
and (2) do not represent views of the way a language description should
look that are held by any well-known linguistic theorists and we shall
decline to investigate these possibilities. The two possibilities repre-
sented by (3) and (4) will, however, continue to concern us. The choice
of a particular descriptive framework should not be a matter of the
investigator’s caprice: rather, it should be dictated by the nature of the
data he is analysing. The linguist sets himself the goal of describing a
language or some part of it by making true statements of the greatest
possible generality, without thereby implying any falsehoods. Since,
as we have noted, the sets of statements represented by the ‘hands’ in
diagrams (1)–(4) are not necessarily identical, the individual statements
which comprise these sets may quite possibly differ in generality. For
example, the individual statements in the set represented by the down-
ward ‘hand’ in diagram (1) may possibly be necessarily much less
general, i.e. much more specific, than the individual statements in the
set represented by the bottom rightmost ‘hand’ in diagram (4). The
very nature of the framework presupposed may force some loss of
generality. The descriptive framework which permits the expression
of all the clearly significant generalizations that hold true about a
language is the optimal framework.

Let us look at a concrete and simplified example of the type of
problem we are facing. Consider for the moment just the English
number-names from one to nine and the corresponding semantic
representations /, . . . , ///////////////. We will temporarily ignore the existence of higher numbers.

A device which has been used in linguistic theory for specifying sets whose members are sequences of identical elements of any length is that of designating the repeated element and marking it with an asterisk. Thus the formula in (5) could be used to generate the infinite set of semantic representations for numbers.

(5) /*

Remembering that we wish to associate each member of the universal set of semantic representations for number concepts with the category NUMBER, we may extend (5) to (6).

(6) NUMBER → /*

For our purposes the arrow sign in (6) may be thought of simply as representing the association we wish to express between NUMBER and each member of the set of semantic representations characterized by the formula /*. A formula such as (6) is known as a rule schema. It expresses what otherwise could only be expressed by postulating an infinite number of rules of the form (7).

(7) NUMBER → /
    NUMBER → //
    NUMBER → ///
    NUMBER → ////// etc.

Rule schema (6) has, as we see, a dual function: it is a device which generates an infinite set of objects (in this case semantic representations) and at the same time associates each of these objects with a linguistic category (in this case the universal category NUMBER). The association between a category and an object (whether it be a semantic representation or not) specified by a rule schema naming that category on the left-hand side of the arrow is conventionally represented in the form of a phrase marker or labelled tree diagram. (8) is an example of a phrase marker derived from rule schema (6). This represents formally the
8 Introduction

association characterized by (6) between the semantic representation /// and the category NUMBER. An infinite number of such tree diagrams can, of course, be derived from rule schema (6) by means of a simple algorithm. This algorithm provides that for any application of a rule or rule schema the symbol on the left-hand side of the arrow be written and joined by downward lines to each of the elements on the right-hand side specified by that particular application of the rule.

We can now associate with each of the semantic representations from / to ////////// the appropriate English phonological form. (Remember that we are temporarily ignoring the existence of higher numbers.) This association can be expressed formally by simply writing the appropriate phonological form beneath a tree diagram such as (8), thereby simultaneously expressing the desired connection between the phonological form and the category NUMBER. An example of such a statement is (9). A statement such as (9) is known as a lexicalization

\[(9)\] 

\[
\begin{array}{c}
\text{NUMBER} \\
/// \\
\text{six}
\end{array}
\]

rule. Phonological sequences such as English one, two, three, ..., nine are known as lexical items. To account for English number-names up to nine we will obviously need nine lexicalization rules such as (9). These will include, for example, those of (10). If English had no higher number

\[(10)\] 

\[
\begin{array}{ccc}
\text{(a) NUMBER} & \text{(b) NUMBER} & \text{(c) NUMBER} \\
/ & / & / \\
\text{one} & \text{two} & \text{three}
\end{array}
\]

expressions than nine we could be content with this description. We have specified the universal set of semantic representations for numbers, associated each such representation with the category NUMBER and provided the necessary rules associating arbitrary phonological forms with semantic representations and the category NUMBER. The nonexistence of any higher number-names is expressed adequately by the nonexistence of any further lexicalization rules. This description, consisting of rule schema (6) and nine lexicalization rules of the form
shown in (10), is roughly of the shape shown in diagram (3). Rule schema (6) and the algorithm for constructing phrase markers fulfil the functions of the boxed ‘device’ and the leftmost ‘hand’ in (3). The description we have provided does not appear to fail to express any obvious generalizations that hold true of English number-names from one to nine. For this limited purpose, then, the framework shown in diagram (3) is quite adequate.

Now let us consider another way of stating the same facts. Another method available within linguistic theory of specifying a set with infinite members is to use a rule that is recursive. In such a rule the symbol used on the left-hand side of the arrow reappears on the right-hand side along with other elements specified by the rule. A single application of a recursive rule generates a set of elements of which the symbol on the left-hand side of the rule is a member. This symbol can be used as a basis for a second application of the same rule, and so on ad infinitum. An example of such a rule is (11).

(11) \text{NUMBER} \rightarrow / \text{NUMBER}

One application of rule (11) and the tree-building algorithm connected with rules produces the phrase marker (12). The lower NUMBER in

(12) \text{NUMBER} \rightarrow \text{NUMBER} / \text{NUMBER}

(12) may now be used as a basis for a second application of rule (11), yielding (13). Clearly rule (11) will not be of any use to us unless we have

(13) \text{NUMBER} \rightarrow \text{NUMBER} / \text{NUMBER} / \text{NUMBER}

some means of stopping it from reapplying ad infinitum. If we could not somehow prevent the reapplication of this rule we would only succeed in generating a single phrase marker of infinite size, whereas we want to generate an infinite set of phrase markers of finite size.
10 I Introduction

To do this we provide for the optional interruption of the cyclic re-
application of (11) after any number of applications. That is, we make
the presence of NUMBER on the right-hand side of the arrow in (11)
optional. The notation conventionally used for this purpose is a pair of
parentheses around the optional element. Thus we substitute (14)
for (11).

(14) NUMBER \rightarrow / (NUMBER)

Rule (14) actually expresses two possibilities: either that the element / is
generated or that the sequence / NUMBER is generated on a particular
application of the rule. Thus rule (14) together with the tree-building
algorithm specifies an infinite set of phrase markers whose members
are as in (15). There are a number of ways in which the universal set of

(15) (a) NUMBER / (b) NUMBER / NUMBER
               /
         / NUMBER
            /
      / NUMBER
         /
    / NUMBER

etc.

semantic representations for numbers can be related to the set of
structures as in (15). One simple and obvious way is to postulate a
convention which reads off the bottom lines of phrase markers such as
those of (15). This convention would correctly relate (15a) to the
semantic representation /, (15b) to the semantic representation //, (15c)
to ///, and so on. Another way is to postulate a set of conventions by
which semantic representations are assigned to phrase markers and
their subparts by the application of the arithmetical operation of
addition. Let us say that if a NUMBER in a phrase marker is connected by
a line to / lower down in the phrase marker and to nothing else, then the
semantic representation we assign to that NUMBER is / . Let us also say
that if a NUMBER in a phrase marker is joined by lines both to a lower /
and to a lower NUMBER, then the semantic representation we assign to
the upper NUMBER is that of the sum of / and the semantic representation