MATHEMATICAL LINGUISTICS

Dale A. Enger
(United States Educational Commission in Korea)

Let me point out, first of all, that the term "mathematical linguistics" if accepted uncritically can be a reflection of an unsophisticated view of the relationship of mathematics and linguistics. There is, I might assert, no such thing as mathematical linguistics, just as there is no such thing as mathematical chemistry or mathematical physics or mathematical psychology (even though one item in my bibliography is titled *A Handbook of Mathematical Psychology.*) I will go even further and state that mathematics really has nothing to contribute to linguistics, at least in the sense that its study adds to our knowledge of this field of study.

Mathematics is, rather, a tool that can be used by workers in any branch of science, including linguistics, for two very general purposes: (1) to quantify data, and (2) provide a formal, unambiguous statement of the theory involved, and to explore implications of the theory by constructing mathematical models which can be studied in a way that the real object of the science can rarely, if ever, be studied.

Part of what has been called mathematical linguistics, then, has been simply quantification of linguistic data, that is, statistical studies of various kinds of linguistic phenomena. These have included studies of the frequency of occurence of lexical items, of phonological units and the ways in which they combine with one another, and of structural items of various kinds. At first glance, this seems to be a very simple kind of use of mathematics; you just count things and find out how many of each there are. But statistics, of course, can be used in very sophisticated ways, not only to tell us how much there is of what we are studying, but also in what proportions, and can reveal subtle relationships that exist between different parts of our data.

By their nature, however, statistical studies deal with data, with things and events in the real world. In terms of our science, this means that statisitical studies are concerned primarily with linguistic performance, and although performance has theoretical implications, such studies are of primary interest to those who work in applied linguistics. Studies of word frequency, of phonological distribution, etc., are extremely useful, for example to textbook writers, language

teachers, etc., but are usually of little import to theoreticians. This holds true, also, for most of the work being done in the field of information theory, which is for the most part a practical study of the communictive power of various kinds of natural and artificial languages for the practical ends of designing efficient communication systems.

I am not, however, I want to assure you, one of those people for whom "practical" and "uninteresting" are virtual synonyms. Many of the statistical studies of linguistic performance are quite interesting, especially Zipf's attempts to equate frequencies of some types of linguistic units to psychobiological factors in humans.

I want to concentrate most of my attention, however, on a discussion of the kinds of mathematics that is used increasingly in discussion of theoretical issues in linguistics. It is, in fact, almost impossible these days to pick up a linguistic journal or text that does not discuss linguistic theory in mathematical terms. The situation is made more difficult for most of us by the fact that many of the terms used could easily be taken as non-mathematical ordinary English—but if we so take it, we are misunderstanding the import of the arguments being presented. When, for example, Chomsky speaks of a language as a "set of sentences," we are missing a great deal if we are not aware that the word "set" is a precise mathematical term subject to the conditions that hold in theory of sets. Or, if we read Hockett's recent attack on Chomsky(Hockett 1969) where he argues that a language is an "ill-defined system" rather than, as Chomsky holds, a "well-defined system," it is essential that we understand that the terms "ill-defined" and "well-defined" are mathematical terms with definite theoretical implications and not simple another way of saying "vague" and "precise."

My point here is simply that it is necessary these days, and increasingly so, for linguists to be literate in the kinds of mathematics being used to argue linguitic thorcy. I do not wish to assert that we all need be mathematicians, able to argue theoretical issues in complex mathematical symbols, but literacy in this field is important, and by literacy I mean at least the ability to recognize a mathematically-derived argument when we see one, and ideally the ability to follow the argument through to its conclusion and make some judgement about its validity.

But before I discuss the particular kinds of mathematics in which we need to become literate, I would like to say a few words about the nature of scientific theories, and the process by which they are developed.

First of all, as we learned in high school science courses, theories begin with observation of linguistics, the date we observe are sentences of a languages, either spoken or written, and

the behavior and attitudes of the language. But it is not enough to have a collection of data. The theory must also have a domain (a mathematical term, by the way, as well as an ordinary English word), a subject, or area, or territory on which it focuses and which it attempts to explain. I call attention to this particularly because the domain of linguistic theory has shifted in recent years without many of us having been clearly aware of it. For those of us who cut our linguistic teeth on Bloomfield and his successors the domain of linguistics was clearly and militantly the language itself, the data was exclusively the utterances of the language. Linguistic theory in the hands of the transformational-generative school, however, takes as its domain not the language itself, but the underlying competence of the ideal speaker/hearer. We are no longer explaining language, but the ideal speaker/hearer, and purporting to predict the behavior of this speaker/hearer and to construct a model (another mathematical term) to explain his competence.

I should point out also that the word "ideal" is of very great importance here, for it is generally true of scientific theories that they deal only indirectly with the rather chaotic real world in which we all live. The laws of physics, for example, explain primarily laboratory phenomena over which physicist has precise control, and do not explain directly naturally occurring events, because such events are affected by many factors which are unpredictable and uncontrollable. This is why the domain of linguisiic theory is an *ideal* speaker/hearer, one who never gets tired, never trips over his tongue, never forgets. If we wish to predict, or explain, actual performance, we need not only a competence theory, but also a performance theory which will take into account the various accidents of the real world that interfere with and prevent the realization of the ideal.

A theory, then, has a domain and begins with observation of data within that domain. The theory, however, does not originate from the data. Theories originate, in some unknown way, in the investigator contemplating or studying or playing around with his data. This is rather a bold assertion, because in essence it amounts to a denial of the validity of the inductive method, whereby investigators contemplating their data are drawn necessarily to certain conclusions, or explanations, or theories, and an assertion that theories come by inspiration and are not necessarily true.[1]

This latter statement, that theories are not necessarily true, is one that we can all accept, for we know that theories must be checked against data in the real world. And this is

[1] For a detailed discussion of this point of view and some convincing argumenst againts the validity of inductive method see Karl Popper's *The Logic of Scientific Discovery*.

precisely where mathematics becomes essential, for a theory is essentially an extended argument, usually one that begins with a series of highly abstract propositions (S→NP+VP, for example, which as a term in an argument might be read "If S, then NP+VP") that are assumed to be true and which yield eventually other propositions of a less abstract nature (I+am+a+student, for example) that can be checked against phenomena in the real world. Arguments, however, are highly technical and if they are to be valid must be cast in clear, concise, unambiguous language, that is, in the language of mathematics. Furthermore, if one wishes to explore the implications and ramifications of a theory it is often necessary to construct a mathematical model of the thing or event under study rather than attempting to study that thing itself. This is clearly what physicists do, for example, in studying atomic and sub-atomic particles. These particles obviously cannot be handled or manipulated by the physicist directly. Instead he constructs a model and checks it against the observed behavior of these particles. This is also what transfomational grammarians do: the competence of a speaker of a language cannot be studied directly, so the transformationalist constructs a model of that competence, a grammar of the language, which is at the same time a theory of the language and of the underlying competence of the speaker/hearer. We might note at this point that such a grammar in the Chomskyian formulation is a theory in the sense defined above: it begins with a series of highly abstract propositions (S→NP+VP, etc.) which eventually yield other propositions (strings of phonetic matrices with an associated semantic interpretation and structural description) which can be checked agains phenomena (real utterances and speakers' reactions to them) in the real world.

Because such a grammar is essentially a theory in this sense, it is basically a mathematical formulation. At this point, I would like to outline in very general terms the kinds of mathematics involved in transformational-generative theory. I cannot, in a paper of this nature teach you mathematics, but I hope at least to show you its relevance to linguistic theory.

First.[2] because it is the basis of all mathematics, and basic especially to theories of all types, is *symbolic logic*, which is concerned primarily with the truth value of propositions, a proposition being simply any statement which can be said to be either true or false. A proposition may be either a single such statement, or a composite composites being made up of various statements and a small set of connectives. If "p" and "q" are any two single propositions, then they may be joined as composites in the following manner:

---

[2] Elementary formulations of set theory, relations, functions and orders may be found in Lipchutz, *Set Theory and Related Topics.*

$\sim p$	read as "not p"	negation
$p \wedge q$	read as "p and q"	conjunction
$p \vee q$	read as "p or q"	inclusive disjunction
p/q	read as "p or q but not both"	exclusive disjunction
$p \rightarrow q$	read as "if p then q"	conditional
$p \leftrightarrow q$	read as "p if and only if q"	biconditional

Some propositions are always true, or are *tautologies*: for example, $p \wedge p$. Some propositions are always false, or are contradiction: for example $p \leftrightarrow p$. In transformational-generative grammar each rule of the grammar may be thought of as either a simple or composite proposition which is either true or false. The rule S → NP+VP, for example, is the proposition "a sentence is a NP and a VP." A rule with options bracketed on the right hand side is equivalent to a proposition with exclusive disjunction, a context-restricted rule is equivalent to a composite proposition with a conditional connective, etc. The rules, or propositions of a TG grammar, however, seem to differ from those of most theoretical arguments in that many, if not most, of the constituent propositions cannot be proven true or false in themselves, but only in the light of later modifying propositions. The proposition S → NP+VP, for example, cannot be judged true or false in and of itself, but only in light of later propositions that may delete NP or change its position. Such a chain of rules is probably best thought of as a single composite proposition with a single truth value.

The second type of mathematics we need to know something about is Set Theory. The basic concept here, of course, is that of the "set," which unfortunately has no precise definition, but corresponds roughly to the notion of a collection or aggregation of things. The notions of set membership, of subsets, of the null or empty set, however, are precisely defined and are applicable to linguistic theory, as are the concepts of set union, complementation, intersection, and inclusion. In fact, these concepts are more commonly used in the formulation of grammars than those of symbolic logic, though it makes little difference which is used, since there is an isomorphism between the algebra of propositions (symbolic logic) and set theory, statements in set theory being translatable into statements in the algebra of propositions and vice versa: to say that set A is included in set B ($A \supset B$), for example, is to say that for all elements x, if x is included in set A, then x is included in set B ((Ax) if $(x \in A)$ then $(x \in B)$), which is a proposition to which a truth value can be assigned. Likewise, the other concepts of set theory have their correspondences in symbolic logic: set union corresponds to inclusive disjunction, intersection to conjunction, the null set to contradiction, set

complementation to a negative proposition, etc.

There seem to be two reasons, however, for preferring formulations in set theory to those in symbolic longic. First, as can be seen from the above example formulations in set theory are more concise, and second, certain very complex relationships are more easily defined and developed in terms of set theory than in terms of symbolic logic. These include the concepts of Orders, Relations, and Functions.

The concept of ordering, of course, is familar to use from the discussions in TG grammar on the ordering of rules in both syntax and phonology. All that need be said about orders at present is that orders are usually defined in terms of ordered pairs, where one element is designated as the first element and the other as the second element, and that a set may be either partially or totally ordered, a totally ordered set being one in which all elements of the set are ordered pairwise in respect to all others, and a partially ordered set one in which some (at least one) element is not ordered in respect to some (at least one) other element. The set of integers 1, 2, 3, 4, 5....., for example, is a totally ordered set because it includes the ordered pairs (1,2), (1,3) (2,3), (1,4), (2,4), (3,4)..... The rules of a generative phonology on the other hand can be shown to be only partially ordered since some rules will not be ordered in respect to others.

The concept of orders, furthermore, is basic to relations, since relations are defined between sets where the members of the sets are ordered pairs. In common terms, we can see that this is true if we consider family relations, for example, the relation "is father of." If we say x is the father of y, we are not only stating a relationship, but we have ordered x and y, so that x is the first element and y is the second. If the relationship is a true one (relations are essentially propositional in nature and can be assigned a truth value) then obviously the order cannot be reversed. If x is the father of y, then y is not the father of x! The concept of relations, of course, is extremely important in linguistics, since many of the concept we deal with are relational in nature. When we say that a noun phrase is the subject of the sentence, for example, we are stating a relationship: noun phrase stands in tne relation "is subject of" to the sentence. This is true also when we speak of predicates, objects of verbs, objects of prepositions, etc.

Function may also be thought of ordered pairs since a function is defined as the assignment to each element in a set A a unique element of a set B (the words "each" and "unique" are important here). Functions are sometimes called "mappings"or "transformations," particularly when regarded as transforming an element of one set into an element of

antoher rather than merely assigning it a value. When regarded in these terms it might be possible to regard a TG grammar essentially as a function: phrase structure rules map S into deep structures: transformational rules map deep structures onto surface structures; and phonological rules map surface structures onto phonetic matrices. This holds true in a strict sense, however, only if all deep structures are mapped onto surface structures (recall the "each" in the definition) and if deep structures and surface structures map onto only one surface structure and phonetic matrix, respectively (recall the "unique" in the definition).

Finally, we need to consider briefly the nature of models. Briefly, a model is a relation that preserves isomorphism, an isomorphism being a one-to-one relationship between sets or objects so that you can talk about them as if they were identical and can regard whatever is done in the model system as having been done in the system that you are studying (actually what you are doing here is making a prediction, for you must check and see that the result of your manipulations with the model system check with the real word system). The concept of models has been significant in the theory of linguistics in two ways, first because it has been necessary to construct models of various types of grammars in order to explore the implications of grammars of various degrees power. These models of grammars are called finite automata, and correspond to grammars of various types: a Turing machine (a machine only in an abstract sense) corresponds to a grammar with an unrestricted rewrite system: a linear-bounded automata corresonds to a context-senstitive phrase structure grammar; a push-down storage automata corresponds to a context-free phrase structure grammar, etc. It is through study of these abstract mathematical devices that it has deen determined that a context-free phrase structure grammar is most suitable for generating sentences of a natural language. The concept of models is also of singnificance, as I have already mentioned, because a TG grammar purports to be a model of the ideal speaker/hearer of a language, which means that the rules of such a grammar must be thought of as corresponding one-to-one with something in the mind of such a speaker/hearer. Of course, as pointed out above, if such a grammar is a good model it will make predictions about the behavior of the speaker which can be checked empirically. But at this point, to hark back to the beginning of my discussion, we must keep in mind that the domain of our study is not simply the sentences of the language we are studying, but also the underlying competence of the speaker, and therefore our empirical check must not only determine that the grammar has produces of the language but that the interpretation of these speaker's understanding of them.

If this can be done, it will have been shown not only that languages themselves can

be described by mathematical processes, but also that mentalistic phenomena—if we take competence to be such a phenomenon—can be approached through mathematical means, which is, I suppose, the reason that Chomsky's articles on mathematical linguistics have appeared in a text not on mathematical linguistics but in one devoted to mathematical psychology.

BIBLIOGRAPHY

1. Basic mathematical texts for linguists

Lipschutz, *Set Theory and Related Topics*

Christian, R. R., *Introduction to Logic and Sets*

Kemeney, Snell, and Thompson, *Introduction to Logic and Sets*

Stoll, *Sets, Logic, and Axiomatic Theories*

2. On the philosophy of science

Popper, Karl, *The Logic of Scientific Discovery*

3. Mathematical studies by linguists

Chomsky, Noam, and George A. Miller, "Introduction to the Formal Analysis of Natural Languages," in Luce, Bush, and Galanter, *Handbook of Mathematical Psychology.*

Chomsky, Noam and George A. Miller, "Finitary Models of Language Users," in *Ibid.*

Herdan, Gustav, *Type-Token Mathematics: A Textbook of Mathematical Linguistics.* Also, see bibliographies in M. Ivic *Thends in Linguistics*, pp.212-242.