Some Observations on the Foundation of Linguistics

William Labov, University of Pennsylvania

1. Fundamental agreements and disagreements in linguistics,

Linguistics is a relatively unified field of study, compared to many others. Over the course of the long history of linguistic analysis, dating from Indian grammarians in the 4th century B.C., we find emerging a high degree of consensus on the fundamental categories like sentence, phrase, noun, verb, vowel and consonant. There is also a wide range of agreement on fundamental principles, and the concepts of language structure enunciated by Saussure (1922) and Bloomfield (1933) in the early part of this century are introduced to students in all elementary texts. It is agreed that linguistics are not interested in a given corpus of linguistic data in itself, but rather in the rules, system and faculty of language that enable speakers to produce that corpus. It is agreed that that language is a system of abstract categories that are mutually defined by their oppositions; that sentences cannot be understood by combination of the meaning of words, but only through the progressive combination of smaller phrases to produce a tree-like structure of immediate constituents.

At the same time, we can observe a profound division in the foundations of our discipline, that corresponds quite closely to the traditional philosophical opposition of idealism and materialism. (In the linguistic literature, this opposition is sometimes described as "mentalism" or "rationalism" vs. "empiricism"). The idealist approach is exemplified by generative grammar, as originated and developed by Chomsky (1957, 1965, 1981), and various other treatments that would account for the same data by parallel methods: generalized phrase structure grammar, lexical-functional grammar, and others. The materialist position is exemplified by the practice current in phonetics, historical linguistics, and dialectology. The principles of this position have been developed most explicitly in sociolinguistics, and in particular in the quantitative study of linguistic variation, which will be the basis of the discussion to follow.

The two approaches, idealist and materialist, differ sharply in their approaches to the foundations of the field: definition of language itself, the methods for gathering data and analyzing it, and the goals of linguistic activity.

2. The definition of the field.

Linguistics does not have a privileged claim upon language as an object of study; there are many other disciplines that examine it – psychology, speech pathology, rhetoric, literary studies, semiotics, and so on. As noted above, linguistics focuses upon abstract language structure, and in particular the phonology, morphology and syntax of the language rather than the vocabulary, idiom or style. Within this area, there are twp opposing answers to the question, "What is language?" The idealist conception is that language is a property of the individual, a species-specific and genetically inherited capacity to form rules of a particular type, relatively isolated from other activities of the human intelligence. The materialistic conception is that language is a property of the speech community, an instrument of social communication that evolves gradually and continuously throughout human history, in response to a variety of human needs and activities.

Subjective vs. objective sources of data. The terms "idealism" and "materialism" can be seen to be most appropriate in relation to the definitions of data involved. The idealist position is that the data of

linguistics consists of speakers' opinions about how they should speak: judgments of grammatically or acceptability that they make about sentences that are presented to them. (These judgments are sometimes referred to as 'intuitions', though at the outset it was clear that the intuitions that actually govern the language are not immediately accessible to direct questioning.) The speaker involved is often the theorist, so that theory and data are simultaneously produced by the same person at the same time.

The materialist approach to the description of language is based on the objective methods of observation and experiment. Subjective judgments are considered a useful and even indispensable guide to forming hypotheses about language structure, but they cannot be taken as evidence to resolve conflicting views. The idealist response is that these objective observations of speech production are a form of 'data flux' which are not directly related to the grammar of the language at all.

The opposition between the idealist and materialist position on data resources is a long-standing one in linguistics, long antedating generative grammar. Among the leaders of American linguistics in the first half of this century, Boas and Sapir favored the basic strategy of 'ask the informant.' Bloomfield, Harris and Voegelin (1951) distrusted the subjective bias of this procedure, and argued that the invention of the magnetic tape recorder should make it possible to base linguistic work on spontaneous speech production. They thus anticipated the full development of the materialist position, and in particular its materialist base.

Definitions of data: 'performance' vs. 'competence.' The idealist position has more recently been reinforced by a distinction between 'performance' and 'competence.' What is actually said and communicated between people is said to be the product of 'language performance', which is governed by many other factors besides the linguistic faculty, and is profoundly distorted by speaker errors of various kinds. The goal of linguistics is to get at the underlying 'competence' of the speaker, and the study of performance is said to lie outside of linguistics proper. The materialist view is that 'competence' can only be understood through the study of 'performance', and that this dichotomy involves an infinite regress: if there are separate rules of performance to be analyzed, then they must also comprise a 'competence', and then new rules of 'performance' to use them, and so on.

The object of description: the individual or the community. The two approaches to the definition of language differ radically in regard to the unit of description. Since judgments of acceptability differ radically and unpredictably across individuals, it is normal for any disagreement about data to be answered by narrowing the unit of description to the 'dialect' of an individual, usually the theorist. Since each individual derives the rule system from fragmentary data, it is generally held that the community is an inconsistent mixture of consistent individuals.

The materialist position begins with the study of the heterogeneity of the speech community, and reduces this variation to a series of regular quantitative patterns controlled by social factors. Early statements about the speech community emphasized this 'structured heterogeneity' as the fundamental feature of the speech community, maintained by a uniformity of social evaluation. More recently, the uniformity of these variable patterns has been found to be also based on a structural homogeneity. In cities of a million or more population, the basic categories ands rules that define the variables are almost constant across social class, sex and age. This reinforces the position that the fundamental unit of description should be the language of the speech community, and that the speech of an individual can only be understood against this background.

The relation of production to perception. The traditional position of linguistics has been that

grammars should be neutral to production and perception. In the idealist view, if either should be given precedence it would be perception. The judgments of grammaticality that are the input to these theories are closer to perception than production, and it is often pointed out that there are people who cannot speak who appear to understand language very well.

The materialist position is that production is methodologically and epistemologically prior to perception. Though the ability to decode language is an essential component of the language faculty, there appears to be far more variability in the routes used in decoding than in the output of speech production. Consistent productive control of a rule system would appear to the most definitive evidence for the existence of that system in the language faculty of a speaker or a community.

2. Mathematical models

The two approaches to the study of linguistics also differ radically in the mathematical models used for the analysis of data, and the assumptions that lie behind these models.

The fundamental postulates. he fundamental postulate of linguistics, as stated most clearly by Bloomfield in 1926, is that "some utterances are the same," (or to be more exact, we can observe partial morphosemantic identities among utterances). Against this postulate we must place the fundamental observation of phonetics " that no two utterances are the same. No matter how hard we try to duplicate what we have just said, the new utterance will vary in many details from the first.

The contradiction between these two principles is resolved by the recognition that native speakers use and react to categories that are linguistically the same, though they may be phonetically different. These "emic" categories " phoneme, morpheme, syntagmeme " are the fundamental currency of linguistic theory. Different elements within a category are in "free variation".

Discrete vs. probabilistic models The two different approaches to linguistic analysis differ in their treatment of this residual variation. In the idealist approach, there is nothing further to be said once free variation is identified: it is not coherent to say that a form applies more often or less often in a given environment. The materialist approach recognizes that free variation can often be constrained by statements about the probability of application in one environment or the other, and that these quantitative constraints can be used to describe the system of the community and the validity of the rules written.

Algebras and probabilities. The mathematical devices used to model the idealist approach to language are qualitative and algebraic, and closely modeled on the algebra of sets. A given rule of the language exists in only one of three states: it always applies in the defined environment; it never applies; or it applies optionally, in "free variation." In this position, there is no theoretically coherent statement that can be made abut how often a rule applies, or whether it applies more often in one environment or another. In fact, some proponents of the idealist position have argued that it is not possible for a human being to learn to do one thing more than another.

The materialist approach begins with the observed variability characteristic of speech production, and applies to this variation formalisms based on probability theory. To do so, it is necessary to define the envelope of variation in a precise manner, so that the sum of the probabilities of all variants equals one. Multivariate analyses can then be applied to the data, to extract the contribution of each element of the environment to the application of a rule.

Rule schemata. A typical rewrite rule used in linguistics takes the form:

(1) X -> Y/A__B

to be read as *X is replaced by Y when it is preceded by A and followed by B*. The rule may also apply when A alternates with C, or B with D; in this case, rule (1) may be elaborated as

(2) X -> Y/

which is an abbreviation of four rules representing all possible combinations of preceding and following environments. Writing such a rule is equivalent to the statement that the choice of A vs. C is independent of the choice of B vs. D, as far as its effect on the rule is concerned. Much of the formal linguistics concerns the creation of such rule schemata, condensing and collapsing individual rules.

The quantitative analysis carried out under the materialist program can tell us whether this operation is justified, testing the validity pf the assumption of independence of the choices of A vs. B and C vs. D. A multivariate analysis is based on a maximum likelihood model using a logistic model, which assigns a weight to each individual factor A, B, C..instead of each combination of environments can be predicted accurately by assigning weights to each individual factor, then these factors can be said to operate independently, and the reduction of individual rules to a rule schema is justified.

Asymmetry of the two models. The algebras developed in formal linguistics, very often under idealist assumptions, are essential starting points for this mode of analysis. There is therefore a marked asymmetry between the two bodies of linguistic activity: those doing empirical analysis can use the formal, qualitative analyses developed under an idealist program, but not visa-versa. The latter are satisfied to construct rule schema without testing for validity against the data of speech production, while the former are not.

This transition from qualitative to quantitative analysis is a familiar one in the development of science. But the qualitative model of linguistics is not easily displaced. Many forms of linguistic behavior are categorically invariant. Furthermore, the number, variety and complexity of linguistic relations are very great, and it is not likely that a large proportion can be investigated by quantitative means. At present, we do not know the correct balance between the two modes of analysis: how far we can go with unsupported qualitative analysis based on introspection, before the proposals must be confirmed by quantitative studies based on observation and experiment.

3. The goals of linguistics.

Earlier in this century, American linguistics saw the primary task of linguistics as the development of a theory of language description. Chomsky expanded this view to include his own conception of the *explanation* of linguistic relations. Explanation of this type is a process internal to the linguistic data: a number of rules are reduced to a single more complex rule, thus capturing a generalization. The original goal was to write a grammar that would produce all and only all the well formed sentences of the language, in the most generalized or *simplest* form. More recently, generative grammar has retreated from the goal of describing individual languages, and taken as a goal the discovery of the universal principles of linguistic structure that govern such generalizations, together with parameter settings that account for differences between languages.

The materialist position is inherently skeptical of the search for universals in the absolute sense of forms or relations that are found in every language without exception. A more realistic type of cross-linguistic study appears to be the search for statistical generalizations at a more concrete level initiated by Greenberg (19??). The materialist view finds the union of theory and practice in a close link to the procedures and aims of linguistic description, and a further development of the general theory of linguistic description. *Explanation* is the materialist approach is a search for substantive correlates of linguistic behavior in its physical and social substrata.

In the materialist view, an adequate description of language must contain a dynamic and evolutionary perspective. Much of the research in the materialist paradigm is devoted to the study of linguistic change, and in particular linguistic change in process. This perspective includes a *uniformitarian* view: that the conditions that led to change in the past are still operating in changes that we observe today. These uniform conditions include innate and physiological factors as well as general principles of linguistic structure and features of the social setting of the use of language. As noted in section 1, the materialist approach views language as an evolutionary product rather than a sudden, species-specific mutation, and many of the principles of the materialist position have been developed in working out the empirical foundations for a theory of language change (Weinreich, Labov and Herzog 1968).

4. Research on linguistic foundations

The view of linguistics that I have presented here is that of a field unified in certain fundamental assumptions and working tools, but deeply divided in others. What type of research on the foundations of linguistics would be helpful in achieving further unification and accelerating progress in the field?

The search for a stable notation. It seems to me that the first priority should be given to the development of a stable notation for the description of languages. It was pointed out initially that linguistics has such a stable notation for basic categories in phonology and grammar. But as more complex syntactic and phonological relations have been uncovered, their representations in the literature have undergone a series of uncontrolled and fanciful changes. Linguist who are engaged in describing natural languages are ill advised to use such unstable notations. Those who tried to do so in the 1960s now find their work unreadable and irrelevant. But descriptions of languages and dialects that make no attempt to incorporate current issues of abstract grammar are condemned to triviality. The answer must lie in the formulation of a "middle range" notation that will do for syntax and phonology what the International Phonetic Alphabet does for phonetics.

The reliability and validity of introspection. The most obvious hiatus in the foundations of modern linguistics is the absence of a concern for the reliability and validity of the introspective judgments that form the main data base of grammatical research. Much linguistic research must be carried out by the direct elicitation of data, especially in the initial stages of investigation, and in the study of syntactic forms of low frequency.

One approach to resolving the opposition between subjective and objective approaches is embodied in the following "working principles for continued exploration of grammatical judgments:" (Labov 1975:30).

I. The Consensus Principle: if there is no reason to think otherwise, assume that the judgments of any native speaker are characteristic of all speakers of the language.

II. The Experimenter Principle: if there is any disagreement on introspective judgments, the judgments of those who are familiar with the theoretical issues may not be counted as evidence.
III. The Clear Case Principle: disputed judgments should be shown to include at least one consistent pattern in the speech community or be abandoned.

As Newmeyer points out (1983:65) these principles assume that the object of investigation should be the speech community, rather that the individual, an issue which still divides the idealist and materialistic approach. In any case, there are no accepted experimental methods that would resolve disputed judgments. Linguists are building on sand until they can answer basic questions: what are the test-retest reliabilities of judgments of grammatical acceptability? Under what conditions do introspections match speech production? What are the sources of bias? Many hundreds of authors have published articles based on introspective data, but only a half dozen have been concerned with this issue. In Anthropology, on the other hand, we find a long-standing and serious concern with informant accuracy (Freeman, Romney and Freeman 1987).

The perceptual correlates of variation. Most of the work on linguistic variation carried out in the materialist approach is based on a study of speech production. Through production may be the preferred source of information on linguistic knowledge, one cannot decide the relevance of these data to the linguistic system without research on the perceptual correlates of variation. Our basic notions about the form and content of grammars cannot be developed fully until we know more about how listeners absorb, code and represent the vast amount of data on linguistic variation that they receive in the course of every-day life.

Work on these three areas would establish the foundations of linguistics on a more secure base, and if successful, would go far towards bridging the fundamental divisions of the field that I have sketched in these pages.