DAVID SANKOFF

PROBABILITY AND LINGUISTIC VARIATION

Many of the most original and well-known applications of probabilistic models and statistics have been made in fields peripheral to linguistics. We may cite Shannon's information theory [42], Zipf's law [47] and other approaches to word frequency distribution [18], the authorship attribution of the Federalist papers by Mosteller and Wallace [29], Osgood's semantic differential [31], and the psycholinguistic analysis of acoustic phonetics in conjunction with the development of multidimensional scaling [43]. In spite of all this, grammarians in the mainstream of linguistics have accorded, until recently and with few exceptions, a uniformly hostile reception to suggestions that probabilistic concepts deserve some consideration in the study of language. This rejection is usually phrased in terms of the conviction that randomness and indeterminacy play no significant role in the structure of language, particularly with respect to the linguistic competence, or internalized linguistic 'knowledge', of the individual speaker-hearer. The genesis of this attitude is threefold, and we will discuss the three aspects in order of increasing importance.

The first is historical. Since the publication in 1957 of Chomsky's *Syntactic Structures* [6], the prevailing viewpoint in linguistics has been that of the generative grammarian. The book began by destroying two 'straw man' theories of grammar, both of which were naive probabilistic models. While this attack served as a useful motivation for the phrase structure and transformational types of generative grammars, as a byproduct it more generally discredited probability notions in the eyes of many linguists and linguistics students. This rejection of probability has since been reaffirmed by Chomsky, for example in his authoritative *Aspects of the Theory of Syntax* [7], published in 1965.

> . . . though 'probability of a sentence (type)' is clear and well-defined, it is an utterly useless notion. [p. 195].

Despite the dominant influence of Chomsky on a whole generation of

Synthese 37 (1978) 217–238.

linguists, it is not his fiat against probability which explains the profound anti-probabilism of many very innovative scholars and indeed the majority of pre-Chomskian linguists. And neither do other historical reasons, such as the controversy over lexicostatistics [8, 10], a statistical approach to the study of the divergence between related languages.

A second aspect has to do with the methodology of grammatical investigation. Linguists have available a very direct, instantly accessible, and very extensive source of data on the grammar of a language, data which has no real counterpart in other fields of study. The data consist of native speakers' yes–no judgments on such matters as the grammaticality of sentences or constructions, or the equivalence of meaning of two sentences (paraphrase). Analytical methods based on these data have been, and are, far more important than other investigative approaches.

In practice, the native speaker is most frequently none other than the linguist herself or himself, and the data are produced simply through introspection. This leads, of course, to a certain number of problems such as a lack of reproducibility between speakers, but these are considered a small price to pay for the insight obtained about linguistic structure. Since a given linguist will be internally self-consistent these will be no variability in his data and no statistical treatment will be necessary. Indeed, the discrete and unvarying nature of an individual's grammaticality judgments are antithetical to notions of quantitative measurement or repeated trials. Yet this too seems insufficient to explain the exclusion of statistical concepts, given the inter-individual, dialectal and historical variation in language which everyone recognizes. In fact it is the study of this variation which has finally provoked much of the recent probabilistic and statistical work which I will be discussing. Nonetheless, many students of linguistic variation refuse to integrate probabilistic concepts into their theories, and continue to produce discrete, deterministic analyses which account more or less well for the phenomena they study.

The third and most important aspect of anti-probabilism has to do with the structure of language, especially as it is understood by modern linguists. So that, referring to our earlier examples, information theory, with its Markovian model of word or letter order, seems a travesty of the complex co-occurrence constraints of syntax or phonology. Similarly, the co-

occurrence phenomena and the logical relationships of semantics, and the close interdependence of semantic and syntactic considerations are ignored and all but blotted out in spatial representation models, such as that associated with the semantic differential. The point is that linguistic theory is formulated in rather mathematical terms, but the type of mathematics involved is far from probability or statistics. It is related more to such areas as the theory of programming languages, automata theory, logic, graph theory and related disciplines, sometimes subsumed under the name 'applied algebra' [1, 3]. These fields are the mathematical subjects closest in spirit to modern linguistics. Their notations and those of linguistics are frequently the same and their historical development has been intertwined with progress in the study of language. Probabilistic extensions of these theories are recent, difficult and not well-known, perhaps because there has been relatively little empirical motivation. It is significant that the very word 'nondeterministic' has quite a different connotation in these fields from what it does in probability theory.

Thus the best understanding of the lack of probability and statistics in main-stream linguistics is through a consideration of the inapplicability of standard probabilistic models and statistical analyses, at least in their usual formulations, to the algebraic structures conceived by linguists as underlying language. This also suggests how we may go about the statistical study of language in a linguistically acceptable way; namely by constructing probabilistic extensions of existing algebraic linguistic models. We shall discuss a few of these without going into too much detail. Research in this field is still very spotty. In some areas much probability theory has been done, but we have few applications to natural language. In other areas statistical methods have been elaborated and applied to fragments of grammars without much consideration of how these might fit into more global probabilistic models.

It is important to stress how the impetus for this work comes from sociolinguistics and related areas such as urban dialectology and historical linguistics [41]. The important distinction between these trends, and the earlier examples also deriving from areas peripheral to linguistics, is that now the motivation for probabilistic extensions comes from within linguistics, by linguists who feel the need to expand the conceptual framework and formal basis of the field in order to understand the variable phenomena under study.

1. PROBABILISTIC GRAMMARS

We may model a language as a countable set of finite sequences which we call the grammatical sentences of the language. The terms in the sequences we call words. Any sequence not in the language is said to be ungrammatical and would be intuitively recognized as such by a hypothetical speaker of the language. A grammar for the language is, roughly speaking, some finite set of machinery, e.g. grammatical rules, whose use, possibly in a recursive manner, can generate all the sentences in the language and none of the ungrammatical sentences.

Classes of grammars can be defined through the types of grammatical rules they contain. A class which is very simply defined, with very restricted types of rules, will generate a restricted class of languages. Such a class, for example the finite-state grammars [3], may be easy to work with mathematically and may have many strong algebraic properties. At the same time, a grammar in this class will not be a very good illustration of any natural language. On the other hand, if we allow very general types of rules in our grammars, such as transformations, the class may contain grammars which are close descriptions of natural language. Unfortunately, what is gained in descriptive adequacy, is lost in mathematical tractability and insight. There will be few mathematical properties of interest which pertain to this class of grammars or their generated languages.

On class of grammars which falls in between these two extremes is the class of context-free grammars. These are relatively simple to define and can be easily and profitably investigated from a mathematical point of view. Moreover, it is quite feasible to construct context-free grammars which simulate reasonably well many of the syntactic properties of natural languages.

The original idea that context-free grammars are suitable structures on which to experiment with probability measures occured independently, to a number of researchers, dating back to Grenander in 1967 [13]. (See also the paper by Klein in 1965 [19].) Some more recent references are [38,44].

A context-free grammar consists of a finite number of rules, for example

(1) $S \rightarrow ASAR$

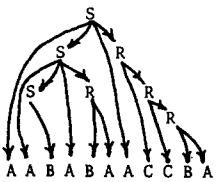
(2) $S \rightarrow B$

(3) $R \rightarrow BA$

(4) $R \rightarrow CR$

where there is exactly one term on the left-hand side (LHS) of each arrow, and all these terms belong to a set of *non-terminal symbols*, in this case $\{S, R\}$. These terms may also appear on the right-hand side, as may the *terminal symbols* or words $\{A, B, C\}$. One non-terminal symbol, S, is specially distinguished, and each sentence is generated from S by a sequence of steps called a derivation, as follows. The first step is to write down "S". Choose any rule in the grammar which *rewrites* S, i.e. for which S is on the LHS, e.g. rule (1). The second step consists of erasing the S and replacing it by the $ASAR$ on the RHS of (1). At each successive step we examine the string of symbols produced by the previous step to see if there are any non-terminal symbols. Each such non-terminal symbol is erased and replaced by the RHS of some appropriate rule. For example $ASAR$ might be rewritten first using (1) and (4) as $AASARACR$, then by (2), (3) and (4) as $AABABAACCR$, and finally by (3) as $AABABAACCBA$. There being no non-terminal symbols left to rewrite, we call the resulting string a *sentence* in the language generated by the grammar. The recursivity of (1) and (4) guarantee an infinite number of sentences in the language. Examples of the sentences are B, $ABABA$, $ABACCCCBA$, etc.

Many programming languages can be generated by context-free grammars, and many properties of natural language grammars can be expressed in context-free form. A sentence derivation can be represented as a labelled directed tree with root at S. For example



Any vertex labeled by a non-terminal symbol has outgoing edges to vertices corresponding to the terms on the RHS of a rewrite rule. In addition these trees are oriented in the plane so that any pairs of vertices X, Y where X and Y are not on the same directed path have a left-right relation; X is to the left

of Y or vice-versa. This is determined by the order of the symbols on the RHS of the rules used to generate the sentence. These trees bear a strong resemblance to the parsing diagrams taught to schoolchildren, and for good reason – they are essentially the same thing. The non-terminal symbols stand for such thing as Noun, Verb, Article, Preposition, Verb Phrase, Noun Phrase etc. As in natural languages, context-free grammars may well generate ambiguous sentences, that is a single sentence can be derived in two or more ways whose tree representations are different. For example, the sentence $ABAB$ can be generated two ways, and is hence ambiguous, by the grammar

$$S \to SU$$

$$S \to A$$

$$S \to ABA$$

$$U \to B$$

$$U \to BAB,$$

as can be seen by inspecting the two derivations:

As already mentioned, there is a rather obvious way of probabilizing context-free grammars. For each non-terminal symbol X, let $r_1, r_2, \ldots, r_k$ be the rules in the grammar which rewrite X. To rule r_i we assign probability p_i in such a way that $p_1 + p_2 + \ldots + p_k = 1$. Similarly for the other non-terminal symbols. This procedure defines a probability for each derivational tree, namely the product of the probabilities of all the rules used in constructing the tree. (This is based on the independent choice of the rules chosen for rewriting each occurrence of each non-terminal symbol.) The probability of a sentence in a language is the sum of the probabilities of all the derivational trees (only one if the grammar, or at least the sentence, is unambiguous) which can be associated with it. Derivation of a sentence becomes a stochastic process which satisfies the definition of a multitype

Galton-Watson branching process, familiar as a model for certain types of population growth [28].

Of what interest are these probabilistic grammars? First, as abstract entities in mathematical linguistics, they have a number of noteworthy properties. For example, they provide a comfortable conceptional framework for dealing with, or dispensing with, infinite derivations, including those that do not produce finite length sentences. Though the possibility of such derivations and such sentences does not disturb algebraic linguists, it does seem advantageous to speak of the probability that a sentence derivation will be finite and will thus produce a sentence. Indeed, there is a very simple condition on the rule probabilities p_i which will ensure that the probability of deriving a finite sentence is equal to one [36]. Another topic of interest is the entropy of probabilistic grammars and their languages. This represents an important improvement over the syntactic naivete of the original Shannon theory insofar as it was meant to be pertinent to natural language.

More interesting than these mathematical considerations, from the linguistic viewpoint, is that probabilistic context-free grammars can serve as primitive models of linguistic variation. Consider a context-free grammar whose rules are fixed, but whose rule probabilities are variable. By setting the probabilities of recursive rules very low, we can ensure that derivations will tend to terminate after the application of a very few rewrite operations. This will result in short 'choppy' sentences. By increasing these same probabilities almost to the point where infinite derivations can occur with non-zero probability, we obtain very long sentences (e.g. with many subordinate clauses). Thus by varying rule probabilities only within an otherwise fixed grammar, we are in effect modelling stylistic variation. This was discussed at greater length by Klein [19].

Now consider two probabilistic grammars on the same set of terminal and non-terminal symbols, representing two different but related dialects, or two historical stages of a single language. Make a larger context-free grammar through the union of the two sets of rules. By varying the rule probabilities according to some parameter representing time or location, we can model the gradual geographic or historical transformation of one speech variety into another.

In much the same way we can model an infant's acquisition of language, starting with one or two non-recursive rules, adding new rules at low levels of

probability, increasing the stock of terminal symbols, and finally introducing recursive rules and adjusting the probabilities to adult levels. This was the basic idea of the language acquisition model introduced by Peizer and Olmsted [32], and forms the basis of the empirical studies of Suppes [45] on child language.

Perhaps the most substantial and significant empirical study making use of probabilistic grammars is that carried out by the Germanistische Seminar at Heidelberg, led by Klein and Dittmar [15, 16]. These researchers undertook a full-scale investigation of the variability in the 'pidgin' German spoken by immigrant workers in Germany. They constructed a sample of 48 Spanish and Italian workers, and developed and applied special techniques for eliciting natural discourse in the code they were studying. From each interview they sampled a hundred sentences which they parsed according to a detailed context-free grammar. From this they estimated the rule probabilities for each speaker. In examining these probabilities they were able to point out some rather interesting phenomena. For each non-terminal symbol, it was possible to construct a linear ordering of the rewrite rules, generally speaking from less complicated to more complicated, such that each speaker could be assigned, on the basis of his probabilities, a position on a linear scale. The remarkable result, which confirms the validity of the analysis, was that the set of rules associated with each of the different non-terminal symbols, all resulted in much the same ordering of the speakers. The researchers then noticed a relation between the relative position of each speaker and his age of arrival in Germany. The scope and depth of this project, together with the detailed and interesting results which emerged, make it one of the most innovative and successful empirical studies based on a probabilistic linguistic model.

The empirical studies we have mentioned all give rise to a certain class of mathematical and statistical problems, which we may call *grammatical inference*. Given a sample of sentences from the language generated by a context-free grammar, what can we infer about the grammar and its probabilities? One version of this problem is in terms of a fixed, given grammar, where only the probabilities need be inferred. (For an unambiguous grammar, this problem is trivial, since all sentences can be parsed in a unique manner and the rules generating them counted.) We have investigated this problem, making use of the properties of the Galton-Watson branching

process already mentioned [36], and exploring the mathematical consequences, for the inference problems, of the left-right structure of derivational trees [38]. This leads as well to the question of which grammars are *inferable* at all, i.e. for which grammars is the statistical information contained in large samples of sentences, sufficient to estimate rule probabilities? Inferability, though motivated by probabilistic considerations, is defined and is a desirable property for context-free grammars in general.

The more difficult problem of inferring the whole grammar, and not just its probabilities, from samples of the language, has been studied in depth by Horning [17]. He has actually computer implemented an integrated algorithmic and Bayesian approach, based on sequential sampling, and capable of making such inference for restricted, but non-trivial, examples.

2. VARIABLE RULES

However convenient it is mathematically to work with probabilistic context-free grammars, and despite the significant empirical contributions of Suppes and of the Heidelberg workers using these grammars, they are not really satisfactory models of natural languages. For example, many of the phonological properties of natural language cannot be accounted for in terms of context-free rules, but require *context-sensitive* rules. The grammar may allow for a term R to be rewritten as BA, but only when the R is found in a string with a T on the immediate left and a U on the immediate right. This is written

$$(5) \qquad R \rightarrow BA/T\_\_U$$

Another rule sight be

$$(6) \qquad R \rightarrow U/\_\_XY$$

which allows the possibility of rewriting R as U when followed by XY. Then the string $TTRRXYZ$ can be rewritten as $TTRUXYZ$ using rule (6), and then as $TTBAUXYZ$ using the rule (5). It is clear that we could not have applied rule (5) unless we had previously applied rule (6), and we had to have applied (6) to the second and not the first R in the original string. This strong dependence between different steps in the derivation clearly contrasts with the independence characteristic of context-free grammars, and we can no

longer make the even stronger assumption of probabilistic independence which is made in that case. Another difference is the blurring of the distinction between terminal and non-terminal symbols. True, there may be some symbols in the grammar which may and must be rewritten and are always found in contexts which permit their rewriting, and there may be some which can never be rewritten. For the most part, however, the eventuality of a symbol's being rewritten, or alternatively, of appearing in the final string or sentence, depends on where the symbol is to be found in the string at the various stages of the particular derivation.

A probabilization of a context-sensitive grammar must include a rule, either deterministic or probabilistic, for selecting at each stage of a derivation which of the symbols of the current string should be examined next for possible rewriting. Different selection rules usually give very different results, since as we have already seen, the application of a rule at one point in the string may very well make possible the application of a second rule to another term, or block the application of a rule which was previously applicable, by altering the context of the term to be rewritten.

The context-sensitive grammars necessary in phonology must be generalized still further to account for many important syntactic properties of language. Because the local structure of one part of a sentence may depend on aspects of another part remote from the first, we need rules in which the environment contains an arbitrarily long *dummy* term such as

$$R \rightarrow BA/T \_\_ xU$$

This is to be read: and R is to be rewritten as BA if it is immediately preceded by T and if there is a U in the string anywhere to the right of the R. The x thus represents any sub-string of length 0, 1, 2, ... These very general rules are called transformations. Many, if not most, transformations are obligatory. That is they must apply whenever the LHS term is in an appropriate context. But it is the non-obligatory, or optional rules which will interest us, from the probabilistic viewpoint. Because of their generality, grammars and the class of grammars containing transformations have none of the strong mathematical properties of context-free grammars. It is, of course, theoretically possible to probabilize these grammars, but the probabilistic structures would be of necessity very complicated and not endowed with any nice properties. Much of formal linguistic research has been involved with the search for universal

restrictions on transformations which would lend transformational grammars additional structure, but any definitive algebraic theory is not likely to be forthcoming in the near future, much less a probabilized version.

Fortunately, if we wish to discuss inference, even without a well-defined global theory, there is a very acceptable way of doing so. Linguists can and usually do analyse a specific syntactic or phonological problem in terms of a fragment grammar containing only those rules perceived to be directly pertinent to the problem. It is generally possible to parse sentences (which are the data to be analyzed) to the extent that those rules within the fragment grammars which have been applied, can be identified, much as with unambiguous context-free grammars. However, unlike context-free grammars it is no longer a very interesting inference problem to estimate the probability of each optional rule's application. The reason for this is that context-sensitive and transformational rules usually occur in 'bundles' of similar rules which are expressed in a condensed version, as for example

$$(7) \qquad R \rightarrow BA \Big/ \begin{Bmatrix} T \\ U \\ V \end{Bmatrix} \_\_ \begin{Bmatrix} K \\ L \\ M \\ N \end{Bmatrix} x \begin{Bmatrix} G \\ H \\ I \end{Bmatrix}$$

This is an abbreviation for the 36 rules

$$R \rightarrow BA \,/\, T \_\_ KxG$$
$$R \rightarrow BA \,/\, U \_\_ KxG$$
$$R \rightarrow BA \,/\, T \_\_ LxH$$
etc.

where the x is any (dummy) string, as above, and the braces enclose the various possible elements of the context which permit the rule to apply. It seems counter-intuitive, both linguistically and statistically, to simply estimate 36 separate probabilities, especially when, as often happens, the presence of one particular symbol in the appropriate position of the context seems to account for a systematic raising, or systematic lowering, of the rule application probability. Thus we are lead to a statistical analysis of the effects of the various symbols in the environment of an optional context-sensitive or transformational rule on rule application probability. And it quickly becomes apparent that even those optional rules within a transformational grammar

which have a context-free form are context-sensitive in a probabilistic sense. That is, though the LHS is always susceptible to being transformed into the RHS, independent of context, in fact this transformation occurs such more frequently in some contexts than others. In empirical linguistics, probabilistic rules of this type were introduced by Labov [21], who called them *variable rules*.

The inference question for variable rules looks very much as if analysis of variance (ANOVA) would be the appropriate methodology. Consider each of the m positions in the context or environment of a rule (in rule (7), $m = 3$) as one of the 'ways' in an m-way design, and measure relative frequencies of rule application under all possible combinations of symbols. Then ANOVA will produce estimates of the effect of each symbol on the rule application probability.

Linguistic data, however, turn out to be unsuitable for the particular estimation procedure used in ANOVA. It is the nature of strings produced by grammars that not all possible combinations of symbols are generated (i.e. are grammatical), so that many large regions of the data table are blank, and these will be distributed in a very unbalanced way. Those cells which linguistically may have entries, in practice tend to have very different number of replications, i.e. rule-eligible strings, ranging from zero to hundreds. Thus in almost every case, the routine application of ANOVA is not appropriate for linguistic variation data. There are various ways of 'patching up' ANOVA to render it applicable [9], but these are very approximate and do not conserve the desirable statistical properties of ANOVA.

Instead of this, we and others [5,22] have adopted a more fundamental though more difficult approach to the estimation of the effects of different factors on rule application probabilities. This involves positing an additive model much as in ANOVA,

$$(8) \qquad X_{ij\ldots k} = \alpha_i + \beta_j + \ldots + \gamma_k$$

where α_i is the effect of the ith factor in the first 'way' or dimension or factor group, β_j is the effect of the jth factor in the second group, etc. In Labov's original model

$$(9) \qquad X_{ij\ldots k} = P_{ij\ldots k}$$

the probability that the rule applies in the context containing factors

$i, j, \ldots, k$. Alternative models proposed by Cedergren and Sankoff require either that

$$(10) \qquad X_{ij\ldots k} = \log P_{ij\ldots k}$$

or that

$$(11) \qquad X_{ij\ldots k} = \log(1 - P_{ij\ldots k})$$

The model which seems currently the most appropriate has

$$(12) \qquad X_{ij\ldots k} = \log \frac{P_{ij\ldots k}}{1 - P_{ij\ldots k}}$$

These last three models have interesting probabilistic interpretations. They may be rewritten as

$$(10a) \qquad P_{ij\ldots k} = P_0 \times P_i \times P_j \times \ldots \times P_k$$

$$(11a) \qquad (1 - P_{ij\ldots k}) = (1 - P_0) \times (1 - P_i) \times (1 - P_j) \times \ldots \times (1 - P_k)$$

$$(12a) \qquad \frac{P_{ij\ldots k}}{1 - P_{ij\ldots k}} = \frac{P_0}{1 - P_0} \times \frac{P_i}{1 - P_i} \times \frac{P_j}{1 - P_j} \times \ldots \times \frac{P_k}{1 - P_k}$$

respectively. In each case, the P_j, say, which is directly related to β_j, can be interpreted as representing the probability that the rule would apply in the (unrealizable) situation where only the factor j is present in the environment, and there is no i factor or k factor, etc. The parameter P_0 represents a sort of starting point for the model, analogous to the overall mean μ in ANOVA. In addition each of these models needs one mathematical constraint per factor group in order to assure uniqueness, analogous to the constraint in ANOVA that row means sum to zero.

Then (10a) embodies the hypothesis that the binomial trial consisting of the decision to apply or not to apply the rule may be decomposed into a number of decisions whether or not to apply it in the environment containing only the factor i, in the environment containing only the factor j, etc. These individual trials are assumed to be probabilistically independent, and the rule is assumed to apply in the presence of all the factors if and only if it applies in *all* the individual trials. Model (11a) is similar except that the rule applies in the presence of all the factors if and only if it applies in *at least one of* the individual trials. These two models are mirror images in that one assumes that

the event that the rule applies requires it apply in all the individual trials, while the other assumes the contrary for the event that the rule does not apply. Unfortunately, both assumptions cannot be made at the same time and it is often difficult to decide which of the two models best fits the data and to interpret the implications of the choice. Model (12a) on the other hand, treats application probabilities and non-application probabilities symmetrically and incorporates both assumptions by restricting independence [33]. It assumes that the individual trials are independent, but that they all result in identical decisions, either to apply the rule or not. This can be seen in the conditional probability statements

$$P_{ij\ldots k} = \frac{P_0 \times P_i \times P_j \times \ldots \times P_k}{P_0 \times P_i \times P_j \times \ldots \times P_k + (1-P_0) \times (1-P_i) \times (1-P_j) \times \ldots \times (1-P_k)}$$

$$(1-P_{ij\ldots k}) = \frac{(1-P_0) \times (1-P_i) \times (1-P_j) \times \ldots \times (1-P_k)}{P_0 \times P_i \times P_j \times \ldots \times P_k + (1-P_0) \times (1-P_i) \times (1-P_j) \times \ldots \times (1-P_k)}$$

which follow from (12a).

The inference problem for these models is as follows. Given that an eligible environment containing factors $i, j, \ldots, k$ has occurred $N_{ij\ldots k}$ times in the language materials being analyzed, and that the rule was applied only $A_{ij\ldots k}$ times, and the corresponding data for many other combinations of factors, how do we estimate the probabilities P_i, P_j, etc., which combine to form the $P_{ij\ldots k}$? To answer this we have recourse to the fundamental principle of maximum likelihood. The likelihood function is set up based on elementary properties of the binomial distribution, but its maximisation requires careful and efficient application of techniques of non-linear programming.

Again perhaps the most important aspect of this development has to do with linguistic variation. Once we are calculating various linguistic influences on rule application it is an easy step to enter into the program factors having to do with style, conversational context, sociodemographic characteristics of speakers, etc. Thus variable rules have had their greatest importance in the study of dialectal variation, linguistic change, acquisition, and the social stratification of language.

Cedergren [4] studied a number of phonological rules in Panamanian Spanish, and with the help of a variable rule analysis was able to detail the intermingling of rural-urban dialectal distinctions with social class conditioning of phonological variation. By grouping speakers in her sample according to age, she was able to show systematic tendencies for rule probabilities to change over time.

In Labov's study of the English copula [22], he was able to show, using estimates derived in a variable rule analysis, how copula deletion in Black English occurred in a quantitatively parallel way to copula contraction in Standard English, depending on the syntactic environment of the variable. At the same time, the analysis enabled him to separate out the effects of phonological conditioning on the operation of the contraction and deletion rules, and to show that they contrasted sharply.

Guy [14] undertook a detailed study of individual variation in final stop deletion in English and showed, using large amounts of data, that the probabilistic parameters associated with the different linguistic factors influencing the rule tended to be stable from individual to individual though there were certain systematic differences among speakers of different dialects.

Laberge [20] used a variable rule analysis to investigate the syntactic, semantic, pragmatic and social factors influencing the use of pronouns *on*, *tu*, *vous* and *ils* to represent an indeterminate referent by speakers of Montreal French, as well as other a number of other variables. This massive study, based on same 20,000 tokens of various pronouns in 150 hours of recorded speech, gave a detailed picture of the ongoing social differentiation of the pronominal system over time, and how this is related to the discourse function of the indeterminate pronouns.

Another large-scale study is that of Labov and Labov [24] on the acquisition of the syntax of questions by their daughter Jessie. They documented 25,000 questions of which some 3,000 were used in a detailed variable rule analysis of the inversion transformation in WH-questions. With this method they carried out a longitudinal study of the tokens by separately analysing the tokens grouped together from each time period. They were able to detect and measure a number of subtle phenomena. The most striking of these is the transition from an early phrase-structure rule producing sentences directly in 'inverted' form, to the situation existing in adult grammars where the phrase-structure rules produce sentences of the same form as assertions

but including a question marker, and which must be inverted by a rule in the transformational component of the grammar.

Other interesting examples of variable rule analyses include Lavandera's thesis on the syntax of conditional sentences in Argentinan Spanish [25], Lefebvre's work on plural marking in Quechua [26], Boyarin's study of phonological variability in Aramaic assessed through orthographic variation in texts [2], and Greenblatt's investigation of metric infractions in 17^{th} century English verse [12]. See also [30, 40].

3. THE LEXICON

Another area I would like to discuss is the relationship between words and meanings. Again we have probabilistic developments based on simple linguistic models of an algebraic nature, as well as the beginning of statistical analyses in the same spirit.

The first generalization of the naive concept of a one-to-one relationship between a word and its meaning is to allow for several words to take on the same meaning and for a single word to take on several meanings. We will base our discussion on one such generalization which is largely motivated by the theory of semantic fields [35].

Let W be a finite set corresponding to the words of a language and M a finite set (usually much smaller than W) representing the primitive components of meaning, e.g. M might contain the syntactic and semantic markers such as 'noun', 'abstract', 'concrete', 'inanimate', 'human', etc. We may define a dictionary as a set of logical implications of form $w \Rightarrow d$, one for each word w in W, and d is the disjunction of several clauses, e.g.

$$d = c_1 \text{ or } c_2 \text{ or } c_3$$

where each clause c is the conjunctions of markers in M, e.g.

$$c = m_1 \text{ and } m_4, \text{ and } m_8 \text{ and } m_9.$$

The set S of all clauses occuring in the dictionary is of course partially ordered by implication. That is if $c_1 \Rightarrow c_2$ and c_1 and c_2 are different, we cannot have that $c_2 \Rightarrow c_1$. This partial order is related to another partial order on some subsets of W as follows. We define the semantic field $S(w)$ of a word w as the set of clauses which can be expressed by w, namely all the

clauses c_1, c_2, etc. in its definition, together with any clause which implies one of these. For example, if $w \Rightarrow c_1$ or c_2, and $c_2 = m_6$ and m_7, then c_2 would be in $S(w)$ as would m_6 and m_7 and m_1, m_6 and m_7 and m_2, m_1 and m_2 and m_6 and m_7, etc.

For any clause c in S, we define its lexical representation set $L(c)$ to be the set of words which can take on the meaning c, namely just those words which have c in their semantic fields. The different $L(c)$ are of course partially ordered by set inclusion. Under certain conditions pertaining to how economically the set M has been chosen, this partial order turns out to be isomorphic with the implicational structure on S. In any case, if $c_1 \Rightarrow c_2$, then $L(c_2)$ is a subset of (or perhaps equal to) $L(c_1)$.

This definition of a dictionary is of course oversimplified (many words require quantifiers in their definition, the relationship between inclusion of the lexical representation sets and the implication of meanings does not always hold, presence or absence of markers is not always sufficient to capture certain distinctions, etc.); However it does capture many important properties of the lexicon. Beside multiple meaning and synonomy, it represents very well the distinction between genetic and specific meaning, and process such as anaphora where less specific words, such as pronouns, are used when the meaning conveyed could be expressed, or has previously been expressed, by a more specific lexical item.

To probabilize this structure, we require a bivariate distribution P on the Cartesian product $W \times S$ with the condition that $P(w, c)$ is greater than zero if and only if meaning c is in the semantic field $S(w)$ of word w, i.e. if and only if w is in $L(c)$, the lexical representation set of c. The quantity $P(w, c)$ represents the marginal probability over some underlying model of speech production (involving probabilistic grammars such as we have discussed, plus some probabilistic way of generating the subject matter of sentences), that the word w is used and the meaning c is expressed. Word frequency lists [18] thus become the study of $\Sigma_c P(w, c)$, for all w. At least one meaning frequency list has also been published [11], and this deals with $\Sigma_w P(w, c)$, for all c. Actual estimates of quantities analogous to $P(w, c)/\Sigma_w P(w, c)$, the conditional probability that w will be used for a given meaning c have been made by Labov [23] for certain w pertaining to cups, bowls and the like. Conversely, Lehrer [27] has estimated quantities similar to $P(w, c)/\Sigma_c P(w, c)$, the conditional probability that meaning c will be conveyed by the

given word w, over the semantic domain of containers. Both these studies were motivated by questions of linguistic variability both between speakers and even for one given speaker, and the present model can be considered as a common mathematical framework in which both may be inserted.

In a recent study [39], we undertook to explore the utility of this model by applying it to certain specific domains within the lexicon. For example, we considered a number of meanings closely related to the notion of 'dwelling', and the words used to express them in Montreal French. We distinguished five meanings, (which in English could be expressed by remain, stay, live, dwell and cohabit, respectively) and four verbs which could take on various of these meanings. This was done with the help of syntactic and semantic criteria, which also facilitated the classification of some 1200 tokens occuring in recorded conversations, in order to estimate the various $P(w, c)$. It was also possible to establish that the values of $P(w, c)$ varied systematically according to the socioeconomic level of the speaker.

A similar study was carried out for meanings related to 'work', 'job', 'task' etc. Here not only the $P(w, c)$ varied according to the speaker, but also certain details of the meaning structure S. A third area investigated, within the conceptual framework we have been discussing, included words and meanings pertaining in the general notion of 'thing', 'object', 'entity'.

On the theoretical level, we can show how our simple way of looking at word-meaning relationships can model some interesting properties of natural languages. Consider a hypothetical probabilized dictionary at a time zero. If we allow small random changes in the $P(w, c)$ over time, for those $P(w, c) > 0$, as well as the occasional possibility for a word to taken on a new meaning, i.e. for $P(w, c)$ to change from zero to a positive value, and the reverse possibility when a $P(w, c)$ becomes zero, we have defined a stochastic model for the evolution of the lexicon over time. Computer simulation experiments give the following results [34, 37]. First, independent of the initial values of the $P(w, c)$, if the experiment is allowed to run long enough, a calculation of $\Sigma_c P(w, c)$ always produces the same results. If these sums, which are overall word frequencies, are ranked in decreasing order, we find that they describe a negative exponential curve as a function of this rank order, which is none other than Zipf's law of word frequency [47]. Second, when we measure the average proportion of meanings with the same lexical representation sets at different points in time, we find that this proportion

declines approximately exponentially with time, which is consistent with Swadesh's theory of glottochronology [46].

4. CONCLUSION

The algebraic structures of formal linguistics incorporate, in a fundamental way, mechanisms of choice which allow a small grammar to generate a diverse and uncountable language. This is illustrated by the choice of rule to rewrite a non-terminal symbol in a context-free grammar, by the choice of whether or not to apply a given phonological or syntactic rule in an eligible environment, and by the choice of which synonym is chosen to express a given meaning.

The fact that grammatical structures incorporate choice as a basic building block means that they accept probabilization in a very natural way, mathematically speaking. This may be seen either in terms of naive probability theory, or through the identification of appropriate σ-algebras over sets of sentences, suitable for the imposition of probability measures. This latter approach is most rewarding in the case of context-free grammars where the relationship with the theory of branching processes is well-established.

The impetus for actually carrying out the probabilization project in linguistics, and the *a posteriori* justification for this work, from the utility standpoint, comes from the field of linguistic variation. Once probabilized, an algebraic structure which is originally categorical, discrete and abstract, suddenly becomes flexible, variable and directly usable as a model of behaviour subject to statistical inference.

In concluding, it is perhaps important to discuss the interpretation of the probabilistic component of linguistic models. It is clear that the regularities and tendencies modeled by these probabilities are of different kinds and come from different sources. On the phonological level, they may often be considered direct consequences of natural articulatory processes, and as such are simply quantitative generalizations of discrete, non-variable (contraction, liason, etc.) phenomena traditionally studied by phonologists and phoneticians. On the semantic level, probabilities are clearly dependent on the topic of the discourse, on the style of conversation, the relationship among speakers and other psychological and sociological factors, and should be considered analytical abstractions rather than components of language.

The same is true though generally to a lesser extent, for syntax. Thus while purely linguistic processes are partly responsible for the nature of the probabilities in question, it would be a mistake to neglect the fact that they are in large part rooted, in a very complex way, in sociological and interactional regularities. Indeed, variation theory is in large part the study of to what extent these probabilities are intrinsic to language as a system, and how extrinsic considerations impinge.

Centre de recherches mathématiques
Université de Montreal

BIBLIOGRAPHY

[1] Birkhoff, G. and Bartee, T. C., 1970, *Modern Applied Algebra*, McGraw-Hill, New York.
[2] Boyarin, D., 1976, 'Variable rules in philology', paper presented to the Linguistic Society of America.
[3] Brainerd, B., 1971, *Introduction to the Mathematics of Language Study*, Elsevier, New York.
[4] Cedergren, H. J., 1973, *Interplay of Social and Linguistic Factors in Panama*, Ph. D. dissertation, Cornell University.
[5] Cedergren, H. J. and Sankoff, D., 1974, 'Variable Rules: Performance as a Statistical Reflection of Competence', *Language* 50, 333–355.
[6] Chomsky, N., 1957, *Syntactic Structures*, Mouton, The Hague.
[7] Chomsky, N., 1975, *Aspects of the Theory of Syntax*, MIT Press, Cambridge, Mass.
[8] Chretien, C. D., 1962, 'The Mathematical Models of Glottochronology', *Language* 38, 11–37.
[9] Cox, D. R., 1970, *Analysis of Binary Data*, Methuen, London.
[10] Dobson, A. J., Kruskal, J. B. Sankoff, D. and Savage, L. J., 1972, 'The Mathematics of Glottochronology Revisted,' *Anthropological Linguistics* 14, 205–212.
[11] Eaton, H. S., 1940, *Semantic Frequency List for English, French, German & Spanish*, University of Chicago Press, Chicago.
[12] Greenblatt, D. L., 1975, 'Variable Rules and Literary Style,' Second International Conference on Computers in the Humanities, Los Angeles, University of Southern California.
[13] Grenander, U., 1967, *Syntax-controlled Probabilities*, Technical Report, Brown University, Division of Applied Mathematics.
[14] Guy, G. R., 1975, 'Variation in the Group and the Individual: the Case of Final Stop Deletion', *Pennsylvania Working Papers on Linguistic Change and Variation 1, No. 4*, Philadelphia, U.S. Regional Survey.
[15] Heidelberger Forschungsprojekt 'Pidgin-Deutsch', 1978, 'The Acquisition of

German Syntax by Foreign Migrant Workers', in *Linguistic Variation*, ed. by D. Sankoff, Academic Press, New York.

[16] Heidelberger Forschungsprojekt 'Pidgin-Deutsch', 1975, 'Zur Sprache ausländischer Arbeiter: Syntaktische Analysen und Aspekte des kommunikative Verhaltens', *Zeitschrift für Literaturwissenschaft und Linguistik* 18, 78–121.

[17] Horning, J. J., 1969, *A Study of Grammatical Inference*, Technical Report No. CS 139, Computer Science Department, Stanford University.

[18] Juilland, A., Brodin, D. and Davidovitch, C., 1970, *Frequency Dictionary of French Words*, Mouton, The Hague.

[19] Klein, S., 1965, 'Control of Style with Generative Grammar, *Language* 41, 619–631.

[20] Laberge, S., 1977, *Étude de la variation des pronoms sujets définis et indéfinis dans le français parlé à Montréal*, Ph.D. dissertation, Université de Montréal.

[21] Labov, W., 1969, 'Contraction, Deletion and Inherent Variability of the English Copula', *Language* 45, 715–62.

[22] Labov, W., 1972, *Language in the Inner City*, University of Pennsylvania Press, Philadelphia.

[23] Labov, W., 1973, 'The Meanings of Words and their Boundaries', in *New Ways of Analyzing Variation in English*, ed. by C. -J. N. Bailey and R. W. Shuy, 340–373, Georgetown University Press, Washington, D.C.

[24] Labov, W. and Labov, T., 1976, 'Learning the Syntax of Questions', in *Proceedings of the Conference on Psychology of Language*, Stirling, Scotland, in press.

[25] Lavandera, B. R., 1975, *Linguistic Structure and Sociolinguistic Conditioning in the Use of Verbal Tense in si-Clauses (Buenos Aires Spanish)*, Ph. D. dissertation, University of Pennsylvania.

[26] Lefebvre, C., 1975, *Plural Agreement in Cuzco Quechua: Some Aspects of Variation*, Ph. D. dissertation, University of California, Berkeley.

[27] Lehrer, A., 1970, 'Indeterminacy in Semantic Description', *Glossa* 4, 87–110.

[28] Mode, C. J., 1971, *Multitype Branching Processes*, Elsevier, New York.

[29] Mosteller, F. and Wallace, D. L., 1964, *The Federalist*, Addison-Wesley, Reading.

[30] Naro, A. J. and Lemle, M., 1976, 'Syntactic Diffusion', in *Papers from the Parasession on Diachronic Syntax*, Chicago Linguistics Society.

[31] Osgood, C. E. Suci, G. J. and Tannenbaum, P. H., 1957, *The Measurement of Meaning*, Urbana.

[32] Peizer, D. B. and Olmsted, D. L., 1969, 'Modules of Grammar Acquisition', *Language* 45, 60–96.

[33] Rousseau, P., 1978, 'Advances in Variable Rule Methodology', in *Linguistic Variation*, ed. by D. Sankoff, Academic Press, New York.

[34] Sankoff, D., 1969, 'Simulation of Word-Meaning Stochastic Processes', International Conference on Computational Linguistics, Stockholm: KVAL, Preprint 49.

[35] Sankoff, D., 1971, 'Dictionary Structure and Probability Measures', *Information and Control* 19, 104–113.

[36] Sankoff, D., 1971, 'Branching Processes with Terminal Types: Application to Context-Free Grammars', *Journal of Applied Probability* 8, 233–240.

[37] Sankoff, D. 1972, 'Lexical Replacement Process', *Computer Studies in the Humanities and Verbal Behavior* 3, 208–212.

[38] Sankoff, D., 1972, 'Context-Free Grammars and Nonnegative Matrices', *Linear Algebra and its Application* 5, 277–281.

[39] Sankoff, D., Thibault, P. and Bérubé, H., 1978, 'Semantic Field Variability', in *Linguistic Variation*, ed. by D. Sankoff, Academic Press, New York.

[40] Sankoff, G., 1973, 'Above and Beyond Phonology in Variable Rules', in *New Ways of Analyzing Variation in English*, ed. by C. -J. N. Bailey and R. W. Shuy, 44–61, Georgetown University Press, Washington, D.C.

[41] Sankoff, G., 1974, 'A Quantitative Paradigm for the Study of Communicative Competence', in *Explorations in the Ethnography of Speaking*, ed. by R. Bauman & J. Sherzer, 18–49, Cambridge University Press.

[42] Shannon, C. E. and Weaver, W., 1949, *The Mathematical Theory of Communication*, Urbana.

[43] Shepard, R. N., 1972, 'Psychological Representation of Speech Sounds', in *Human Communication: a Unified View*, ed. by E. E. David & P. B. Denes, 67–113, McGraw Hill, New York.

[44] Soule, S., 1974, 'Entropies of Probabilistic Grammars', *Information and Control* 25, 57–74.

[45] Suppes, P., 1970, 'Probabilistic Grammars for Natural Languages', *Synthese* 22, 95–116.

[46] Swadesh, M., 1952, 'Lexicostatistic Dating of Prehistoric Ethnic Contracts', *Proceedings of the American Philosophical Society* 96, 452–463.

[47] Zipf, G. K. 1945, 'The Meaning-Frequency Relationship of Words', *Journal of General Psychology* 33, 251–256.