DOCUMENT RESUME

ED 041 288 AL 002 499

AUTHOR	Suppes, Patrick
TITLE	Probabilistic Grammars for Natural Languages. Psychology Series.
INSTITUTION	Stanford Univ., Calif. Inst. for Mathematical Studies in Social Science.
SPONS AGENCY	National Science Foundation, Washington, D.C.
REPORT NO	TR-154
PUB DATE	15 May 70
NOTE	34p.
EDRS PRICE	EDRS Price MF-$0.25 HC-$1.80
DESCRIPTORS	*Child Language, Evaluation Criteria, *Grammar, *Linguistic Theory, *Mathematical Linguistics, Nominals, Phrase Structure, Probability, Psycholinguistics, *Statistical Studies

ABSTRACT

The purpose of this paper is to define the framework within which empirical investigations of probabilistic grammars can take place and to sketch how this attack can be made. The full presentation of empirical results will be left to other papers. In the detailed empirical work, the author has depended on the collaboration of E. Gammon and A. Moskowitz, and draws on joint work for examples in subsequent sections. Section II presents a simple example of a probabilistic grammar to illustrate the methodology without complications. Section III indicates how such ideas may be applied to the spoken speech of a young child. Because of the difficulties and complexities of working with actual speech, the fourth section illustrates some of the results obtained when the apparatus of analysis is applied to a much simpler corpus, a first-grade reader. The results of an empirical sort in this paper are all preliminary in nature. (Author/AMM)

PROBABILISTIC GRAMMARS FOR NATURAL LANGUAGES

by

Patrick Suppes

TECHNICAL REPORT NO. 154

May 15, 1970

PSYCHOLOGY SERIES

INSTITUTE FOR MATHEMATICAL STUDIES IN THE SOCIAL SCIENCES
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

Probabilistic Grammars for Natural Languages*

Patrick Suppes

1. Introduction

Although a fully adequate grammar for a substantial portion of any natural language does not exist, a vigorous and controversial discussion of how to choose among several competing grammars has already developed. On occasion, criteria of simplicity have been suggested as systematic scientific criteria for selection. The absence of such systematic criteria of simplicity in other domains of science inevitably raises doubts about the feasibility of such criteria for the selection of a grammar. Although some informal and intuitive discussion of simplicity is often included in the selection of theories or models in physics or in other branches of science, there is no serious systematic literature on problems of measuring simplicity. Nor is there any systematic literature in which criteria of simplicity are used in a substantive fashion to select from among several theories. There are many reasons for this, but perhaps the most pressing one is that the use of more obviously objective criteria leaves little room for the addition of further criteria of simplicity. The central thesis of this paper is that objective probabilistic criteria of a standard scientific sort may be used to select a grammar.

Certainly the general idea of looking at the distribution of linguistic types in a given corpus is not new. Everyone is familiar with the remarkable

*This research has been supported in part by the National Science Foundation under Grant NSFGJ-443X. The report will appear as an article in a forthcoming issue of Synthese.

agreement of Zipf's law with the distribution of word frequencies in almost any substantial sample of a natural language. The empirical agreement of these distributions with Zipf's law is not in dispute, although a large and controversial literature is concerned with the most appropriate assumptions of a qualitative and elementary kind from which to derive the law. While there is, I believe, general agreement about the approximate empirical adequacy of Zipf's law, no one claims that a probabilistic account of the frequency distribution of words in a corpus is anything like an ultimate account of how the words are used or why they are used when they are. In the same sense, in the discussion here of probabilistic grammars, I do not claim that the frequency distribution of grammatical types provides an ultimate account of how the language is used or for what purpose a given utterance is made. Yet, it does seem correct to claim that the generation of the relative frequencies of utterances is a proper requirement to place on a generative grammar for a corpus.

Because of the importance of this last point, let me expand it. It might be claimed that the relative frequencies of grammatical utterances are no more pertinent to grammar than is the relative frequency of shapes to geometry. No doubt, in one sense such a claim is correct. If we are concerned, on the one hand, simply with the mathematical relation between formal languages and the types of automata that can generate these languages, then there is a full set of mathematical questions for which relative frequencies are not appropriate. In the same way, in standard axiomatizations of geometry, we are concerned only with the representations of the geometry and its invariants, not with questions of actual frequency of distribution of figures in nature. In fact, we all recognize that such questions are foreign to the spirit of either classical or modern geometry. On the other hand, when we deal with the physics of objects in nature there are many aspects of shapes and their frequencies of fundamental importance, ranging from the discussion of the shape of clouds and the reason for their shape to the spatial configuration of large and complex organic molecules like proteins.

From the standpoint of empirical application, one of the more dissatisfying aspects of the purely formal theory of grammars is that no distinction is made between utterances of ordinary length and utterances that are arbitrarily long, for example, of more than 10^{50} words. One of the most obvious and fundamental features of actual spoken speech or written text is the distribution of length of utterance, and the relatively sharp bounds on the complexity of utterances, because of the highly restricted use of embedding or other recursive devices. Not to take account of these facts of utterance length and the limitations on complexity is to ignore two major aspects of actual speech and writing. As we shall see, one of the virtues of a probabilistic grammar is to deal directly with these central features of language.

Still another way of putting the matter is this. In any application of concepts to a complex empirical domain, there is always a degree of uncertainty as to the level of abstraction we should reach for. In mechanics, for example, we do not take account of the color of objects, and it is not taken as a responsibility of mechanics to predict the color of objects. (I refer here to classical mechanics--it could be taken as a responsibility of quantum mechanics.) But ignoring major features of empirical phenomena is in all cases surely a defect and not a virtue. We ignore major features because it is difficult to account for them, not because they are uninteresting or improper subjects for investigation. In the case of grammars, the features of utterance length and utterance complexity seem central; the distribution of these features is of primary importance in understanding the character of actual language use.

A different kind of objection to considering probabilistic grammars at the present stage of inquiry might be the following. It is agreed on all sides that an adequate grammar, in the sense of simply accounting for the grammatical structure of sentences, does not exist for any substantial portion of any natural language. In view of the absence of even one grammar in terms of this criterion, what is the point of imposing a stricter criterion to also account for the relative frequency of utterances? It might be asserted that until at least one adequate grammar exists, there is no need to be concerned with a probabilistic criterion of choice. My answer to such a claim is this. The probabilistic program described in this paper is meant to be supplementary rather than competitive with traditional investigations of grammatical structure. The large and subtle linguistic literature on important features of natural language syntax constitutes an important and permanent body of material. To draw an analogy from meteorology, a probabilistic measure of a grammar's adequacy stands to ordinary linguistic analysis of particular features, such as verb nominalization or negative constructions, in the same relation that dynamical meteorology stands to classical observation of the clouds. While dynamical meteorology can predict the macroscopic movement of fronts, it cannot predict the exact shape of fair-weather cumulus or storm-generated cumulonimbus. Put differently, one objective of a probabilistic grammar is to account for a high percentage of a corpus with a relatively simple grammar and to isolate the deviant cases that need additional analysis and explanation. At the present time, the main tendency in linguistics is to look at the deviant cases and not to concentrate on a quantitative account of that part of a corpus that can be analyzed in relatively simple terms.

Another feature of probabilistic grammars worth noting is that such a grammar can permit the generation of grammatical types that do not occur in a given corpus. It is possible to take a tolerant attitude toward utterances that are on the borderline of grammatical acceptability, as long as the relative frequency of such utterances is low. The point is that the objective of the probabilistic model is not just to give an account of the finite corpus of spoken speech or written text used as a basis for estimating the parameters of the model, but to use the finite corpus as a sample to infer parameter values for a larger, potentially infinite "population" in the standard probabilistic fashion. On occasion, there seems to have been some confusion on this point. It has been seriously suggested more than once that for a finite corpus one could write a grammar by simply having a separate rewrite rule for each terminal sentence. Once a probabilistic grammar is sought, such a proposal is easily ruled out as acceptable. One method of so doing is to apply a standard probabilistic test as to whether genuine probabilities have been observed in a sample. We run a split-half analysis, and it is required that within sampling variation the same estimates be obtained from two randomly selected halves of the corpus.

Another point of confusion among some linguists and philosophers with whom I have discussed the methodology of fitting probabilistic grammars to data is this. It is felt that some sort of legerdemain is involved in estimating the parameters of a probabilistic grammar from the data which it is supposed to predict. At a casual glance it may seem that the predictions should always be good and not too interesting because the parameters are estimated from the very data they are used to predict. But this is to misunderstand the many different ways the game of prediction may be played. It is certainly true that if the number of parameters equals the number of predictions the results are not very interesting. On the other hand, the more the number of predictions exceeds the number of parameters the greater the interest in the predictions of the theory. To convince one linguist of the wide applicability of techniques of estimating parameters from data they predict and also to persuade him that such estimation is not an intellectually dishonest form of science, I pointed out that in studying the motion of the simple mechanical system consisting of the Earth, Moon and Sun, at least 9 position parameters and 9 velocity or momentum parameters as well as mass parameters must be estimated from the data (the actual situation is much more complicated), and everyone agrees that this is "honest" science.

It is hardly possible in this paper to enter into a full-scale analysis and defense of the role of probabilistic and statistical methodology in science. What I have said briefly here can easily be expanded upon; I have tried to deal with some of the issues in a monograph on causality (Suppes, 1970). It is my own conviction that at present the quantitative study of language must almost always be probabilistic in nature. The data simply cannot be handled quantitatively by a deterministic theory. A third confusion of some linguists needs to be mentioned in this connection. The use of a probabilistic grammar in no way entails a commitment to finite Markovian dependencies in the temporal sequence of spoken speech. Two aspects of such grammars make this clear. First, in general such grammars generate a stochastic process that is a chain of infinite order in the terminal vocabulary, not a finite Markov process. Second, the probabilistic parameters are attached directly to the generation of non-terminal strings of syntactic categories. Both of these observations are easy to check in the more technical details of later sections.

The purpose of this paper is to define the framework within which empirical investigations of probabilistic grammars can take place and to sketch how this attack can be made. The full presentation of empirical results will be left to other papers. In the detailed empirical work I have depended on the collaboration of younger colleagues, especially Elizabeth Gammon and Arlene Moskowitz. I draw on our joint work for examples in subsequent sections of this paper. In the next section I give a simple example, indeed, a simple-minded example, of a probabilistic grammar, to illustrate the methodology without complications. In the third section I indicate how such ideas may be applied to the spoken speech of a young child. Because of the difficulties and complexities of working with actual speech, I illustrate in the fourth section some of the results obtained when the apparatus of analysis is applied to a much simpler corpus, a first-grade reader. I emphasize the results of an empirical sort in this paper are all preliminary in nature. The detailed development of the empirical applications is a complicated and involved affair and goes beyond the scope of the work presented here.

2. A simple example

To illustrate the methodology of constructing and testing probabilistic grammars, a simple example is described in detail in this section. It is not meant to be complex enough to fit any actual corpus.

The example is a phrase-structure grammar that can easily be rewritten as a regular grammar. The five syntactic or semantic categories are just V_1, V_2, Adj, PN and N, where V_1 is the class of intransitive verbs, V_2 the class of transitive verbs or two-place predicates, Adj the class of adjectives, PN the class of proper nouns and N the class of common nouns. Additional non-terminal vocabulary consists of the symbols S, NP, VP and AdjP. The set P of production rules consists of the following seven rules plus the rewrite rules for terminal vocabulary belonging to one of the five categories. The probability of using one of the rules is shown on the right. Thus, since Rule 1 is obligatory, the probability of using it is 1. In the generation of any sentence either Rule 2 or Rule 3 must be used. Thus the probabilities α and $1 - \alpha$, which sum to 1, and so forth for the other rules.

Production Rule	Probability
1. $S \rightarrow NP + VP$	1
2. $VP \rightarrow V_1$	$1 - \alpha$
3. $VP \rightarrow V_2 + NP$	α
4. $NP \rightarrow PN$	$1 - \beta$
5. $NP \rightarrow AdjP + N$	β
6. $AdjP \rightarrow AdjP + Adj$	$1 - \gamma$
7. $AdjP \rightarrow Adj$	γ

Thus this probabilistic grammar has three parameters, α, β and γ, and the probability of each grammatical type of sentence can be expressed as a monomial function of the parameters. In particular, if Adj^n is understood to denote a string of n adjectives then the possible grammatical types (infinite in number) all fall under one of the corresponding schemes, with the indicated probability.

Grammatical Type	Probability
1. $PN + V_1$	$(1 - \alpha)(1 - \beta)$
2. $PN + V_2 + PN$	$\alpha(1 - \beta)^2$
3. $\mathrm{Adj}^n + N + V_1$	$(1 - \alpha)\beta(1 - \gamma)^{n-1}\gamma$
4. $PN + V_2\ \mathrm{Adj}^n + N$	$\alpha\beta(1 - \beta)(1 - \gamma)^{n-1}\gamma$
5. $\mathrm{Adj}^n + N + V_2 + PN$	$\alpha\beta(1 - \beta)(1 - \gamma)^{n-1}\gamma$
6. $\mathrm{Adj}^m + N + V_2 + \mathrm{Adj}^n + N$	$\alpha\beta^2(1 - \gamma)^{m+n-2}\gamma^2$

On the hypothesis that this grammar is adequate for the corpus we are studying, each utterance will exemplify one of the grammatical types falling under the six schemes. The empirical relative frequency of each type in the corpus can be used to find a maximum-likelihood estimate of each of the three parameters. Let $x_1, \ldots, x_n$ be the finite sequence of actual utterances. The likelihood function $L(x_1, \ldots, x_n; \alpha, \beta, \gamma)$ is the function that has as its value the probability of obtaining or generating sequence $x_1, \ldots, x_n$ of utterances given parameters α, β, γ. The computation of L assumes the correctness of the probabilistic grammar, and this implies among other things the statistical independence of the grammatical type of utterances, an assumption that is violated in any actual corpus, but probably not too excessively. The maximum-likelihood estimates of α,

β and γ are just those values of $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\gamma}$ that maximize the probability of the observed or generated sequence $x_1, \ldots, x_n$. Let y_1 be the number of occurrences of grammatical type 1, i.e., $PN + V_1$, as given in the above table, let y_2 be the number of occurrences of type 2, i.e., $PN + V_2 + PN$, let $y_{3,n}$ be the number of occurrences of type 3 with a string of n adjectives, and let similar definitions apply for $y_{4,n}$, $y_{5,n}$ and $y_{6,m,n}$. Then on the assumption of statistical independence, the likelihood function can be expressed as:

$$(1)\quad L(x_1,\ldots,x_n;\, \alpha,\beta,\gamma) = [(1-\alpha)(1-\beta)]^{y_1}[\alpha(1-\beta)^2]^{y_2}\prod_{n=1}^{\infty}[1-\alpha)\beta(1-\gamma)^{n-1}\gamma]^{y_{3,n}}\ldots$$

$$\prod_{n=1}^{\infty}\prod_{m=1}^{\infty}[\alpha\beta^2(1-\gamma)^{m+n-2}\gamma^2]^{y_{6,m,n}} .$$

Of course, in any finite corpus the infinite products will always have only a finite number of terms not equal to one. To find $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\gamma}$ as functions of the observed frequencies $y_1, \ldots, y_{6,m,n}$, the standard approach is to take the logarithm of both sides of (1), in order to convert products into sums, and then to take partial derivatives with respect to α, β and γ to find the values that maximize L. The maximum is not changed by taking the log of L, because log is a strictly monotonic increasing function. Letting $\not{L} = \log L$, $y_3 = \Sigma y_{3,n}$, $y_4 = \Sigma y_{4,n}$, $y_5 = \Sigma y_{5,n}$, and $y_6 = \Sigma\Sigma y_{6,m,n}$, we have

$$\frac{\partial \not{L}}{\partial \alpha} = -\frac{y_1 + y_3}{1-\alpha} + \frac{y_2 + y_4 + y_5 + y_6}{\alpha} = 0 ,$$

$$\frac{\partial \not{L}}{\partial \beta} = -\frac{y_1}{1-\beta} - \frac{2y_2}{1-\beta} + \frac{y_3}{\beta} + \frac{y_4 + y_5}{\beta} - \frac{y_4 + y_5}{1-\beta} + \frac{2y_6}{\beta} = 0$$

$$\frac{\partial L}{\partial \gamma} = \frac{y_3 + y_4 + y_5 + y_6}{\gamma} - \left[\frac{y_{3,2} + y_{4,2} + y_{5,2}}{1 - \gamma} + \dots + \frac{(n-1)(y_{3,n}+y_{4,n}+y_{5,n})}{1 - \gamma} + \dots\right]$$

$$- \left[\frac{y_{6,1,1}}{1 - \gamma} + \dots + \frac{(m-n-2)y_{6,m,n}}{1 - \gamma} + \dots\right] = 0 .$$

If we let

$$z_{6,n} = \sum_{m'+n'=n+1}\sum y_{6,m',n'} ,$$

then after solving the above three equations we have as maximum-likelihood estimates:

$$\hat{\alpha} = \frac{y_2 + y_4 + y_5 + y_6}{y_1 + y_2 + y_3 + y_4 + y_5 + y_6}$$

$$\hat{\beta} = \frac{y_3 + y_4 + y_5 + 2y_6}{y_1 + 2y_2 + y_3 + 2y_4 + 2y_5 + 2y_6}$$

$$\hat{\gamma} = \frac{y_3 + y_4 + y_5 + z_6}{\Sigma n(y_{3,n} + y_{4,n} + y_{5,n} + z_{6,n})}$$

As would be expected from the role of γ as a stopping parameter for the addition of adjectives, the maximum-likelihood estimate of γ is just the standard one for the mean of a geometrical distribution.

Having estimated α, β and γ from utterance frequency data, we can then test the goodness of fit of the probabilistic grammar in some standard statistical fashion, using a chi-square or some comparable statistical test. Some numerical results of such tests are reported later in the paper. The criterion for acceptance of the grammar is then just a standard statistical one. To say this is not to imply that standard statistical methods or criteria of testing are without their own conceptual problems. Rather the intention is to emphasize that the selection of a grammar can follow a standard scientific methodology of great power and wide applicability, and methodological arguments meant to be special to linguistics--like the discussions of simplicity--can be dispensed with.

3. Grammar for Adam I

Because of the relative syntactic simplicity and brevity of the spoken utterances of very young children, it is natural to begin attempts to write probabilistic grammars by examining such speech. This section presents some preliminary results for Adam I, a well-known corpus collected by Roger Brown and his associates at Harvard.* Adam was a young boy of about 26 months at the time the speech was recorded. The corpus analyzed by Arlene Moskowitz and me consists of eight hours of recordings extending over a period of some weeks. Our work has been based on the written transcript of the tapes made at Harvard. Accepting for the most part the word and utterance boundaries established in the Harvard transcript, the corpus consists of 6,109 word occurrences with a vocabulary of 673 different words and 3,497 utterances.

Even though the mean utterance length of Adam I is somewhat less than 2.0, there are difficulties in writing a completely adequate probabilistic grammar for the full corpus. An example is considered below.

*Roger Brown has generously made the transcribed records available and given us permission to publish any of our analyses.

To provide, however, a sample of what can be done on a more restricted basis, and in a framework that is fairly close to the simple artificial example considered in the preceding section, I restrict my attention to the noun phrases of Adam I. Noun phrases dominate Adam I, if for no other reason than because the most common single utterance is the single noun. Of the 3,497 utterances, we have classified 936 as single occurrences of nouns. Another 192 are occurrences of two nouns in sequence, 147 adjective followed by noun, and 138 adjectives alone. In a number of other cases, the whole utterance is a simple noun phrase preceded or followed by a one-word rejoinder, vocative or locative.

The following phrase-structure grammar was written for noun phrases of Adam I. There are seven production rules, and the corresponding probabilities are shown on the right. This particular probabilistic model has five free parameters; the sum of the a_i's is one, so the a_i's contribute four parameters to be fitted to the data, and in the case of the b_i's there is just one free parameter.

Production Rule	Probability
1. NP → N	a_1
2. NP → AdjP	a_2
3. NP → AdjP + N	a_3
4. NP → Pro	a_4
5. NP → NP + NP	a_5
6. AdjP → AdjP + Adj	b_1
7. AdjP → Adj	b_2

What is pleasing about these rules and perhaps surprising is that six of them are completely standard. (The one new symbol introduced here is Pro for pronoun; inflection of pronouns has been ignored in the present grammar.) The only slightly nonstandard rule is Rule 5. The main application of this rule is in the production of the noun phrases consisting of a noun followed by a noun, with the first noun being an uninflected possessive modifying the second noun. Examples from the corpus are *Adam horn*, *Adam hat*, *Daddy racket* and *Doctordan circus*.

To give a better approximation to statistical independence in the occurrences of utterances, successive occurrences of the same noun phrase were deleted in the frequency count, and only first occurrences in a run of occurrences were considered in analyzing the data. Using the resulting 2,352 occurrences of noun phrases in the corpus, the maximum-likelihood estimates of the parameters obtained are the following:

Estimated Parameter Values

$a_1 = .7001$	$b_1 = .0599$
$a_2 = .0966$	$b_2 = .9401$
$a_3 = .0072$	
$a_4 = .0787$	
$a_5 = .1174$	

On the basis of remarks already made, the high value of a_1 is not surprising because of the high frequency of occurrences of single nouns in the corpus. It should be noted that the value of a_1 is even higher than the relative frequency of single occurrences of nouns, because the noun-phrase grammar has been written to fit all noun phrases, including those occurring in full sentence context or in conjunction with verbs, etc. Thus in a count of single nouns as noun phrases every occurrence of a single noun as a noun phrase was counted, and as can be seen from the table below, there are 1,580 such single nouns without immediate repetition. The high value of b_2 indicates that there are very few occurrences of successive adjectives, and therefore in almost all cases the adjective phrase was rewritten simply as an adjective (Rule 7).

Comparison of the theoretical frequencies of the probabilistic grammar with the observed frequencies is given in Table 1.

Insert Table 1 about here

Some fairly transparent abbreviations are used in the table in order to reduce its size; as before, N stands for noun, A for adjective, and P for pronoun. From the standpoint of a statistical goodness-of-fit test, the chi-square is still enormous; its value is 355.0 and there are only three net degrees of freedom. Thus by ordinary statistical standards we must reject the fit of the model, but at this stage of the investigation the qualitative comparison of the observed and theoretical frequencies is encouraging. The rank order of the theoretical frequencies for the more frequent types of noun phrases closely matches that of the observed frequencies. The only really serious discrepancy is in the case of the phrases consisting of two nouns, for which the theoretical frequency is substantially less than the observed frequency. It is very possible that a different way of generating the possessives that dominate the occurrences of these two nouns in sequence would improve the prediction.

TABLE 1

Probabilistic Noun-Phrase Grammar for Adam I

Noun Phrase	Observed Frequency	Theoretical Frequency
N	1580	1646.5
A	244	213.6
NN	231	135.4
P	176	185.2
PN	31	15.2
NA	19	17.6
NNN	12	11.1
AA	10	12.8
NAN	8	1.3
AP	6	2.0
PPN	6	.1
ANN	5	1.4
AAN	4	.9
PA	4	2.0
ANA	3	.2
APN	3	.2
AAA	2	.8
APA	2	.0
NPP	2	.1
PAA	2	.1
PAN	2	.1

Summation of the observed and theoretical frequencies will show there is a discrepancy between the two columns. I explicitly note this. It is expected, because the column of theoretical frequencies should also include the classes that were not observed actually occurring in the corpus. The prediction of the sum of these unobserved classes is that they should have a frequency of 109.6, which is slightly less than 5% of the total observed frequency of 2,352.

It is also to be noted that the derivation of the probabilities for each grammatical type of noun phrase used the simplest derivation. For example, in the case of Adj + N the theoretical probability was computed from successive application of Rule 3, followed by Rule 6, followed by Rule 7. It is also apparent that a quite different derivation of this noun phrase can be obtained by using Rule 5. Because of the rather special character of Rule 5, all derivations avoided Rule 5 when possible and only the simplest derivation was used in computing the probabilities. In other words, no account was taken of the ambiguity of the noun phrases. A more exact and sensitive analysis would require a more thorough investigation of this point. It is probable that there would be no substantial improvement in theoretical predictions in the present case, if these matters were taken account of. The reader may also have noted that the theoretical frequencies reflect certain symmetries in the predictions that do not exist in the observed frequencies. For example, the type Pro + Pro + N has an observed frequency of six, and the permutation N + Pro + Pro has an observed frequency of two. This discrepancy could easily be attributed to sampling. The symmetries imposed by the theoretical grammar generated from Rules 1 to 7 are considerable, but they do not introduce symmetries in any strongly disturbing way. Again it is to be emphasized that the symmetries that are somewhat questionable are almost entirely introduced by means of Rule 5. Finally, I note that I have omitted from the list of noun phrases the occurrence of two pronouns in sequence because all cases consisted of the question *Who that?* or *What that?*, and it seemed inappropriate to classify these occurrences as single noun phrases. I hasten to add that some remarks of a similar sort can be made about some of the other classifications. I plan on a subsequent occasion to reanalyze these data with a more careful attention to semantics and on that occasion will enter into a more refined classification of the noun phrases.

It is important for the reader to keep in mind the various qualifications that have been made here. I have no intention of conveying the impression that a definitive result has been obtained. I present the results of Table I as a preliminary indication of what can be achieved by the methods introduced in this paper. Appropriate qualifications and refinements will undoubtedly lead to better and more substantial findings.

I would like now to turn to the full corpus of Adam I. It is possible to write a phrase-structure grammar very much in the spirit of the partial grammar for noun phrases that we have just been examining. However, since approximately as good a fit has been obtained by using a categorial grammar and because such a grammar exhibits a variant of the methodology, I have chosen to discuss the best results I have yet been able to obtain in fitting a categorial grammar to the data of

Adam I. I emphasize at the very beginning that the results are not very good. In view of the many difficulties that seem to stand in the way of improving them, I shall deal rather briefly with the quantitative results.

Some preliminary remarks about categorial grammars will perhaps be useful, because such grammars will probably not be familiar to some readers. The basic ideas originated with the Polish logicians Lesniewski (1929) and Ajdukiewicz (1935). The original developments were aimed not at natural language, but at providing a method of parsing of sentences in a formal language. From a formal standpoint there are things of great beauty about categorial grammars. For example, in the standard approaches there are at most two production rules. Let α and β be any two categories, then we generate an expression using the right-slant operation by the rule

$$\alpha \rightarrow \alpha/\beta, \beta \ ,$$

and we generate an expression using the left-slant operation by the rule

$$\alpha \rightarrow \beta, \beta\backslash\alpha .$$

In addition, the grammars began with two primitive categories, s and n , standing respectively for sentence and noun. A simple sentence like John walks has the following analysis

$$\begin{array}{cc} \underline{\text{John}} & \underline{\text{walks}} \\ n\ , & n\backslash s \end{array} \ .$$

Note that $n\backslash s$ is the derived category of intransitive verbs. The sentence

$$\begin{array}{ccc} \underline{\text{John}} & \underline{\text{loves}} & \underline{\text{Mary}} \\ n & (n\backslash s)/n, & n \end{array}$$

has the analysis indicated. In this case the derived category of transitive verbs is $(n\backslash s)/n$. In a basic paper on categorial grammars, Bar-Hillel, Gaifman and Shamir (1960) showed that the power of categorial grammars is that of context-free grammars. In a number of papers that mention categorial grammars and describe some of their features, the kind of simple examples I have just described are often given, but as far as I know, there has been no large-scale effort to analyze an empirical corpus using such grammars. (For an extensive discussion see Marcus (1967).)

The direct application of standard categorial grammars to Adam I is practically impossible. For example, in the standard formulation the single axiom with which derivations begin is the primitive symbol s, and with this beginning there is no way of accounting for the dominant number of noun-phrase utterances in Adam I. I have reworked the ideas of categorial grammars to generate always from left to right, to have the possibility of incomplete utterances, and to begin derivations from other categories than those of sentencehood. From a formal standpoint, it is known from the paper of Bar-Hillel, Gaifman and Shamir that a single production rule will suffice, but the point here is to introduce not just a single left-right rewrite rule, but actually several rewrite rules in order to try to give a more exact account of the actual corpus of Adam I.

Although it is possible to write a categorial grammar in these terms for Adam I, my efforts to fit this grammar probabilistically to the frequencies of utterances have been notably unsuccessful. I have spent more time than I care to say in this endeavor. It has been for me an instructive lesson on the sharp contrast between writing a grammar for a corpus without regard for utterance frequencies and writing a probabilistic grammar. Because of the clear failure of the temporal categorial grammar to account for the probabilistic features of Adam I, I shall not enter into extensive details here.

The three left-right production rules were

1. $\beta \rightarrow \beta,\beta\backslash\alpha,$
2. $\alpha/\beta \rightarrow \alpha/\beta,\beta,$
3. $\alpha \rightarrow \alpha,\gamma_1,\ldots\gamma_n,$

provided $\alpha,\gamma_1,\ldots,\gamma_n$ cancels to α under the standard two rules given earlier. Each of these three rewrite rules is used with probability t_i, $i = 1,2,3$. Secondly, generalizing the classical single axiom s, any one of an additional 10 categories could begin a derivation; e.g., n (nouns), n/n (adjectives), l/n (locatives), r/n (rejoinders), s\s (adverbs), s/n (transitive verbs), and so forth. After the generation of each category, with probability σ the utterance terminated, and thus a geometrical distribution was imposed on utterance length. In the use of Rewrite Rule 1, the category α needed to be selected; the model restricted the choice to n,s or v (vocatives). Finally, two categories, the primitive <u>poss</u> for possessives and ϵ for the empty set, were replacements used in applying Rewrite Rule 3. The model just described was applied to the 22 types of utterances having a frequency of 20 or more in the corpus. The most important fact about the poor fit was that the theoretical frequencies were smaller than the observed frequencies in all 22 cases. Much of the theoretical probability was assigned to other utterance types, the effect being to spread the theoretical distribution more uniformly over a larger number of utterance types than was the case for the actual distribution. Just to illustrate the situation I cite the data for the three most frequent utterance types, giving first the observed and then the predicted frequency: n = 626, 422.6; s/n,n = 206, 25.6; r = 168, 133.3. Some readers may properly ask why I should report at all this unsatisfactory temporal categorial grammar. Partly it is just my own lingering affection for these grammars, but more, it is the simplicity of developmental sequence these grammars would offer if successful. With a uniform, fixed set of rewrite rules, only two things would change with the maturation of the child: the list of derived categories, and the values of probability parameters. But I currently see no hope of salvaging this approach.

Because it is natural to point a finger at the left-right feature of the rewrite rules, I should also mention that I tried fitting a grammar based on the two standard rewrite rules given earlier, one going to the left and one to the right, but also without any reasonable degree of success.

4. Grammar for a First-grade Reader

As the analysis of the preceding section shows it is not yet possible to give a fully satisfactory account of the grammatical aspects of Adam I. Preliminary indications for a larger corpus of more than twenty hours' recording of a 30-month old girl are of a similar nature. We do not yet understand how to write a probabilistic grammar that will not have significant discrepancies between the grammatical model and the corpus of spoken speech.

Examining the results for Adam I early in 1969 and once again failing to make a significant improvement over the results obtained with Arlene Moskowitz in 1968, I asked myself in a pessimistic moment did there exist any actual corpus of spoken or written speech for which it would be possible to write a probabilistically adequate grammar. Perhaps the most natural place to look for simple and regular utterances is in a set of first-grade readers. Fortunately Elizabeth Gammon (1969) undertook the task of such an analysis as her dissertation topic. With her permission I use some of her data.

Readers who have not tried to write a generative grammar for some sample corpus may think that this sounds like a trivial task in view of the much talked about and often derided simplicity of first-grade readers. Far be it from the case. Gammon's grammar is far too complex to describe in detail here. Perhaps the most surprising general feature it reveals is that first-grade readers have a wider variety of grammatical forms than of vocabulary. Before she undertook the analysis we had expected a few stereotypic grammatical forms to dominate the corpus with the high frequency of their appearance. The facts were quite different. No form had a high frequency, and a better a priori hypothesis would have been that a large number of grammatical types of utterances were approximately uniformly distributed, although this assumption errs in the other direction.

To provide a good statistical test of the probabilistic ideas developed in this paper, the most practical move is to write grammars for parts of utterances rather than whole utterances, as has already been seen in the case of Adam I.

Using Gammon's empirical count for types of noun phrases in the Ginn Pre-Primer (1957), I have written in the spirit of Sections 2 and 3 two grammars for noun phrases. In the first one the number of parameters is 5. Four of the 7 rules are also used in the NP grammar for Adam I given above. The rule NP → Pro is dropped, but replaced by NP → PN, the rule NP → AdjP is dropped and replaced by the rule NP → N + Adj. This rule is of course derivable from the NP rules for Adam I; we just use Rule 5, then Rules 1, 2 and 7. The rule NP → NP + NP of Adam I is dropped, and a new rule to handle the use of definite articles (T) is introduced: NP → T. In summary form, the grammar G_1 is the following.

Noun-Phrase Grammar G_1 for Ginn Pre-Primer

Production Rule	Probability
1. NP → N	a_1
2. NP → AdjP + N	a_2
3. NP → PN	a_3
4. NP → N + Adj	a_4
5. AdjP → AdjP + Adj	b_1
6. AdjP → Adj	b_2
7. AdjP → T	b_3

Using the 528 phrases classified as noun-phrases in Dr. Gammon's grammar, we obtain the following maximum-likelihood estimates of the parameters of G_1.

$$\begin{array}{ll} a_1 = .1383 & b_1 = .2868 \\ a_2 = .3674 & b_2 = .0662 \\ a_3 = .4697 & b_3 = .6471 \\ a_4 = .0246 & \end{array}$$

Using these estimated values of the parameters, we may compute the theoretical frequencies of all the types of noun-phrases actually occurring in the corpus. The Grammar G_1 generates an infinite number of types, but of course almost all of them have very small theoretical frequencies. Observed and theoretical frequencies are given in Table 2.

Insert Table 2 about here

It is apparent at once that Grammar G_1 fits the Ginn data a good deal better than the grammar for Adam I fits Adam's data. The chi-square of 3.4 reflects this fact. Let me be explicit about the chi-square computation. The contribution of each type is simply the square of the difference of the observed and theoretical frequencies divided by the theoretical frequency. Except that when a theoretical frequency is less than 5, frequencies of more than one type are combined.* In the case of G_1, the theoretical frequency 3.7 for Adj + Adj + N was combined with the residual of 5.6,

*The number 5 is not sacred; it provides a good practical rule. When the theoretical frequency is too small, e.g., 1, 2 or 3, the assumptions on which the goodness-of-fit test is based are rather badly violated.

TABLE 2

Prediction of Grammars G_1 and G_2 for Ginn Pre-Primer

Noun Phrase	Observed* Freq.	Theoretical Freq. of G_1	Theoretical Freq. of G_2
PN	248	248.0	248.0
T + N	120	125.5	129.5
N	73	73.0	66.9
T + Adj + N	42	36.0	34.2
T + Adj + Adj + N	14	10.3	9.1
N + Adj	13	13.0	13.0
Adj + N	10	12.8	17.7
Adj + Adj + N	8	3.7	4.7
	528		

*Data from dissertation of Dr. Elizabeth Gammon the sum theoretically assigned by G_1 to all other types of noun phrases generated by G_1 different from those listed in Table 2. This means the chi-square was computed for an aggregate of 8 cells, 5 parameters were estimated from the data, and so there remained 2 net degrees of freedom. The chi-square value of 3.4 is not significant at the .10 level, to use the ordinary statistical idiom, and so we may conclude that we have no reason for rejecting G_1 at the level of grammatical detail it offers. A closer examination of the way the parameters operate does reveal the following. Parameter a_3 is estimated so as to exactly fit the frequency of noun phrases that are proper nouns (PN), and parameter a_4 so as to exactly fit the frequency of the type N + Adj. Each of these parameters uses up a degree of freedom, and so there is not an interesting test of fit for them. The interest centers around the other types, and this may well be taken as a criticism of G_1. Further structural assumptions are needed that reduce the number of parameters, and especially that interlock in a deeper way the probabilities of using the different production rules. In spite of the relatively good fit of G_1, it should be regarded as only a beginning.

It is a familiar fact that two grammars that have different production rules can generate the same language, i.e., the same set of terminal strings. It should also be clear that as probabilistic grammars they need not be equivalent, i.e., they need not make the same theoretical predictions about frequencies of occurrences of utterance-types. These matters may be illustrated by considering a second grammar G_2 for the noun phrases of the Ginn Pre-Primer.

Noun-Phrase Grammar G_2 for Ginn Pre-Primer

	Production Rule	Probability
1.	NP → AjdP + N	a_1
2.	NP → PN	a_2
3.	NP → N + Adj	a_3
4.	AdjP → AdjP + Adj	b_1
5.	AdjP → T	b_2
6.	AdjP → ϵ	b_3

In the sixth production rule of G_2 the symbol ϵ is used, as earlier, for the empty symbol.

The theoretical predictions of G_2 are given in Table 2, and it is apparent that as probabilistic grammars G_1 and G_2 are not equivalent, the fit of G_1 being slightly better than that of G_2, although it is to be noted that G_2 estimates 4 rather than 5 parameters from the data.

The examples that have been given should make clear how a probabilistic criterion can be imposed as an additional objective or behavioral constraint on the acceptability of a grammar. In a subsequent paper I intend to show how the probabilistic viewpoint developed here may be combined with a model-theoretic generative semantics. In this more complex setup the semantic base of an utterance affects the probability of its occurrence and requires a formal extension of the ideas set forth here.

5. Representation Problem for Probabilistic Languages

From what has already been said it should be clear enough that the imposition of a probabilistic generative structure is an additional constraint on a grammar. It is natural to ask if a probabilistic grammar can always be found for a language known merely to have a grammar. Put in this intuitive fashion, it is not clear exactly what question is being asked.

As a preliminary to a precise formulation of the question, an explicit formal characterization of probabilistic grammars is needed. In a fashion familiar from the literature we may define a grammar as a quadruple (V_N, V_T, R, S), where V_N, V_T and R are finite sets, S is a member of V_N, V_N and V_T are disjoint, and R is a set of ordered pairs, whose first members are in V^+, and whose second members are in V^*, where $V = V_N \cup V_T$, V^* is the set of all finite sequences whose terms are elements of V, and V^+ is V^* minus the empty sequence. As usual, it is intended that V_N be the non-terminal and V_T the terminal vocabulary, R the set of productions and S the start symbol. The language L generated by G is defined in the standard manner and will be omitted here.

In the sense of the earlier sections of this paper, we obtain a probabilistic grammar by adding a conditional probability distribution on the set R of productions. Formally we have:

Definition. _A quintuple_ $G = (V_N, V_T, R, S, p)$ _is a probabilistic grammar if and only if_ $G = (V_N, V_T, R, S)$ _is a grammar, and_ p _is a real-valued function defined on_ R _such that_

(i) _for each_ (σ_i, σ_j) _in_ R, $p(\sigma_i, \sigma_j) \geq 0$,

(ii) _for each_ σ_i _in the domain of_ R

$$\sum_{\sigma_j} p(\sigma_i, \sigma_j) = 1 ,$$

where the summation is over the range of R.

Various generalizations of this definition are easily given; for example, it is natural in some contexts to replace the fixed start symbol S by a probability distribution over V_N. But such generalizations will not really affect the essential character of the representation problem as formulated here.

For explicitness, we also need the concept of a probabilistic language, which is just a pair (L, p), where L is a language and p is a probability density defined on L, i.e., for each x in L, $p(x) \geq 0$ and

$$\sum_{x \in L} p(x) = 1 .$$

The first formulation of the representation problem is then this.

Let L _be a language of type_ i (i = 0, 1, 2, 3), _with probability density_ p. _Does there always exist a probabilistic grammar_ G (_of type_ i) _that generates_ (L, p)?

What is meant by generation is apparent. If $x \in L$, $p(x)$ must be the sum of the probabilities of all the derivations of x in G. Ellis (1969) has answered this formulation of the representation problem in the negative for type 2 and type 3 grammars. His example is easy to describe. Let $V_T = \{a\}$, and let $L = \{a^n | n \geq 1\}$. Let $p(a^{n+1}) = \frac{1}{\sqrt{t_n}}$, $n > 0$, where $t_1 = 4$, and t_i = smallest prime such that $t_i > \max(t_{i-1}, 2^{2i})$

for $i > 1$. In addition, set

$$p(a) = 1 - \sum_{n=1}^{\infty} p(a^{n+1}) .$$

The argument depends upon showing that the probabilities assigned to the strings of L by the above characterization cannot all lie in the extensions of the field of rational numbers generated by the finite set of conditional probabilities attached to the finite set of production rules of any context-free grammar.

From the empirically-oriented standpoint of this paper, Ellis' example, while perfectly correct mathematically, is conceptually unsatisfactory, because any finite sample of L drawn according to the density p as described could be described also by a density taking only rational values. Put another way, algebraic examples of Ellis' sort do not settle the representation problem when it is given a clearly statistical formulation. Here is one such formulation. (As a matter of notation, if p is a density on L, p_s is the sample density of a finite random sample drawn from (L, p).)

Let L *be a language of type* i *with probability density* p. *Does there always exist a probabilistic grammar* G *(of type* i*) that generates a density* p' *on* L *such that for every sample* s *of* L *of size less than* N *and with density* p_s *the null hypothesis that* s *is drawn from* (L, p') *would not be rejected?*

I have deliberately imposed a limit N on the size of the sample in order directly to block asymptotic arguments that yield negative results. In referring to the null hypothesis' not being rejected I have in mind using some standard test such as Kolmogorov's and some standard level of significance. The details on this point do not matter here, although a precise solution must be explicit on these matters and also on problems of repeated sampling, fixing the power of the test, etc. My own conjecture is that the statistical formulation of the problem has an affirmative solution for every N, but the positive solutions will often not be conceptually interesting.

A final remark about the density p on L is perhaps needed. Some may be concerned about the single occurrence of many individual utterances even in a large corpus. The entire discussion of the representation problem is easily shifted to the category descriptions of terminal strings as exemplified in earlier sections of this paper, and at this level certainly many grammatical types occur repeatedly.*

*W. C. Watt has called my attention to an article by Harwood (1959), which reports some frequency data for the speech of Australian children, but no probabilistic grammar or other sort of model is proposed or tested. As far as I know, the explicit statistical test of probabilistic grammars, including estimation of parameters, has not been reported prior to the present paper, but given the scattered character of the possibly relevant literature I could just be ignorant of important predecessors to my own work.

References

Ajdukiewicz, K. 'Die Syntaktische Konnexität', *Studia Philosophica* **1** (1935) 1-27.

Bar-Hillel, Y., Gaifman, C., and Shamir, E. 'On Categorial and Phrase Structure Grammars', *Bulletin of the Research Council of Israel, Section F* **9** (1960) 1-16.

Ellis, C. Probabilistic Languages and Automata, Doctoral Dissertation, University of Illinois at Urbana-Champaign, 1969.

Gammon, E. A Syntactical Analysis of Some First-Grade Readers, Doctoral Dissertation, Stanford University, 1969.

Harwood, F. W. 'Quantitative Study of the Speech of Australian Children', *Language and Speech* **2** (1959) 236-271.

Lesniewski, S. Grundzüge Eines Neuen Systems der Grundlagen der Mathematik', *Polska Akademia Nauk, Fundamenta Mathematicae* **14** (1929) 1-81.

Marcus, S. *Algebraic Linguistics; Analytical Models*, Academic Press, New York 1967.

Russell, D., and Ousley, O. *The Pre-Primer Program: My Little Red Story Book, My Little Green Story Book, My Little Blue Story Book*, Ginn, Boston 1957.

Suppes, P. *A Probabilistic Theory of Causality* , *Acta Philosophica Fennica* **24** (1970).