



## A MEASURE OF DELIBERATENES AS AN AID TO THE CONSTRUCTION OF GRAMMARS

Richard Rohwer\*

The construction of a grammar from a given set of terminal symbols and a corpus illustrating their use is not a particularly straightforward, or even well-defined problem. In most any approach to this problem, it is necessary to know how the terminal symbols should be grouped into phrases. To this end, a measure of "deliberateness" or "non-randomness" of phrases is introduced. This measure can be computed directly from the N-Gram statistics of the corpus, and takes into consideration a simple model of the uncertainties in these statistics. It indicates whether the correlation is positive or negative. A high value for this deliberateness measure appears to be a sufficient, but not necessary condition for the phrase to have relevance to the grammar. The measure can also be used to judge the non-randomness of a production rule. It is concluded that this measure, while unable to provide all the information needed to construct a general phrase-structure grammar, provides a substantial subset of this information. The measure is also useful for computing probabilities for arbitrary strings of terminal symbols.

### INTRODUCTION

A grammar (ref 1) is set of rules for generating sequences. Briefly, these are **rewrite rules** which allow particular sequences of **terminal** and **non-terminal** symbols to be rewritten as different sequences. The **language** generated by the grammar is the set of sequences of terminal symbols which can be obtained by arbitrarily many applications of the rewrite rules to a special non-terminal symbol called the **starting symbol**. The grammar is **probabilistic** if each sequence in each rewrite rule is to be chosen with a prescribed probability. The **statistical inference of grammars** is the problem of finding a probabilistic grammar suitable for a given sample of the language. The sample might be a set of strings of symbols, each string having been derived from the starting symbol, or less informatively, a single string consisting of several such strings run together.

It is always possible to invent a grammar to fit a finite sample of a language. It suffices to state one rewrite rule which transcribes the starting symbol directly into the string or strings of the sample. Of course, a "smaller", more structured solution is usually desired. The number of rules in a "good" grammar, and the space required to specify each rule, should both be small compared to the size of the language sample. In principle one can always find all the "good" solutions, however "good" might be defined, because there are a finite number of grammars which are "small" compared to the language sample, and all of these could be enumerated and inspected. (Strictly speaking, there are an infinite number of probabilistic grammars to look at, because the probabilities can assume any of a continuum of values. But there is no need to specify the probabilities to an accuracy beyond the statistical uncertainties inherent in the finite-sized language sample.) Anyway, the finite but astronomical number of grammars which would have to be investigated renders this approach wildly impractical except in the most trivial cases. An heuristic approach is needed whereby simple statistical properties of the sample can be used to guide one to guess "good" production rules.

In this paper a simple statistic is derived which can be used to discover which symbols are most likely to have been "deliberately" clumped together in the language sample, in the sense that the clumping cannot easily be attributed to pure chance.

### PROBABILITY ESTIMATES AND THEIR UNCERTAINTIES

The simplest type of statistic which can be extracted from the language sample is a **1-gram** probability  $P(x)$ , the probability that symbol  $x$  would be selected if a symbol were selected at random from a language sample.  $P(x)$  can be estimated by the usual formula

$$P(x) = N(x) / L \quad (1)$$

where  $N(x)$  is the number of times  $x$  appears in the language sample,

\*Centre for Speech Technology Research, The University of Edinburgh, 80 South Bridge, Edinburgh, EH1 1HN, UK.

and  $L$  is the number of symbols in the sample; ie.,

$$L = \sum_{y \in T} N(y). \quad (2)$$

Here  $T$  is the set of terminal symbols.

This estimate will vary from one sample to another. Let  $p$  be the correct figure for  $P(x)$ ; the probability of drawing  $x$  from an infinite sample of the language. Let  $q=1-p$  be the probability of drawing a symbol other than  $x$  from the infinite sample. Just as with a weighted coin, the binomial distribution determines the probability of drawing a given number of  $x$ 's from a finite sample of size  $L$ . The binomial distribution has mean  $Lp$  and standard deviation  $\sqrt{Lpq}$  (ref 2). If no one symbol has a high probability compared to 1, then for any symbol  $x$ ,  $q$  is nearly 1.  $N(x)$  serves to estimate the mean  $Lp$ , so one can conclude that  $N(x)$  has an uncertainty of roughly  $\sqrt{N(x)}$ . Thus  $P(x)$  has an uncertainty of roughly

$$\delta P(x) = \sqrt{N(x)}/L. \quad (3)$$

The relative uncertainty in  $P(x)$  is

$$\delta P(x)/P(x) = 1/\sqrt{N(x)}. \quad (4)$$

An **n-gram** is a sequence of  $n$  symbols. The probability and uncertainty for an  $n$ -gram can be estimated just as for a 1-gram, except that the number of occurrences of the length- $n$  sequence of interest,  $N(x_1, x_2, \dots, x_n)$  replaces  $N(x)$  in (1), and the total number of length- $n$  sequences  $L-n+1$  replaces  $L$ . If  $L$  is large (as is required to obtain small uncertainties) then  $L$  is a good estimate of  $L-n+1$ .

## CORRELATION

Suppose that the language is governed by the "random" grammar in which the starting symbol is rewritten with 100% probability as a string of  $L$  identical non-terminal symbols, and that this non-terminal symbol is to be rewritten as any single terminal symbol with a fixed probability for each terminal symbol. This grammar would account perfectly well for the 1-gram probabilities derived from the language sample. If in fact this grammar expressed all the structure contained in the language, then the  $n$ -gram probabilities would be related to the 1-gram probabilities in a simple way (ref 3):

$$P(x_1, x_2, \dots, x_n) = P(x_1)P(x_2) \dots P(x_n). \quad (5)$$

If the grammar is less trivial there will be correlations which render this relationship false. The equality will fail more severely for some  $n$ -grams than others, and this information suggests which sequences of terminal symbols should appear in production rules.

The severity of the failures of (5) must be judged with reference to the uncertainties in the probabilities which appear in the equation. By inspecting the derivative of a product one concludes that the relative uncertainty of a product is the sum of the relative uncertainties of the factors. Therefore if the equality in (5) fails by a fraction which is less than the sum of the relative uncertainties in the probabilities involved, there is no evidence for correlation. (For an elementary discussion of error propagation see (ref 4).)

Let us focus on 2-Grams, and recast (5) as

$$\stackrel{\text{def}}{C_{x,y}} = P(x,y)/P(x)P(y) = 1. \quad (6)$$

A convenient measure of the correlation in the digram  $xy$  is provided by  $\log(C_{x,y})$ . Suppose for example that  $C_{x,y}=2$ . This would mean that the digram  $xy$  occurs twice as often as it would in the absence of correlations. Alternatively if  $C_{x,y}=1/2$ , then  $xy$  occurs half as often. In these two cases  $\log(C_{x,y})$  has the same magnitude and opposite sign. When the sign is positive the correlation is positive, and otherwise it is negative.

The relative uncertainty in  $C_{x,y}$  is

$$1/\sqrt{N(x,y)} + 1/\sqrt{N(x)} + 1/\sqrt{N(y)}. \quad (7)$$

Usually the 1-grams which appear within a 2-gram are much more frequent than the 2-gram, so the last two terms can be dropped. The uncertainty in  $C_{x,y}$  can therefore be approximated as

$$\delta C_{x,y} = C_{x,y} / \sqrt{N(x,y)}. \quad (8)$$

Using the first order Taylor expansion for the (base-e) logarithm one estimates the uncertainty in  $\log(C_{x,y})$  as

$$\delta \log C_{x,y} = \delta C_{x,y} / C_{x,y}. \quad (9)$$

From (8) one finds that this conveniently works out to be  $1/\sqrt{N(x,y)}$ . Let  $Q_{x,y}$  be  $\log(C_{x,y})$  expressed in units of its uncertainty:

$$\text{def} \\ Q_{x,y} = \log(C_{x,y}) / \delta \log C_{x,y}, \quad (10)$$

or

$$Q_{x,y} = \log(C_{x,y}) \sqrt{N(x,y)}. \quad (11)$$

A large absolute value for  $Q_{x,y}$  indicates that the production rules will definitely need to be set up in such a way that  $y$  will follow  $x$   $C_{x,y}$  times as often as it does in the "random" language. This will be the case regardless of whether  $P(xy) > P(x)P(y)$  or  $P(xy) < P(x)P(y)$ , and regardless of whether  $C_{x,y}$  is especially large or small. The fact that  $Q$  is essentially a measure of  $C$  in *units of its uncertainty* implies that  $Q$  picks out the cases of most certain relevance.

#### USING THE CORRELATION INFORMATION

There are various ways in which a grammar can be defined to produce a high value of  $Q_{x,y}$  for the string  $xy$ . The most obvious way involves having a nonterminal symbol, say  $A$ , which can be rewritten as  $xy$  with high probability. There must also be other rewrite rules which produce  $A$  with reasonably high probability; otherwise  $xy$  would not have occurred often enough to have the statistical significance implied by a large  $Q$ . A strongly negative  $Q_{x,y}$  can be arranged if  $A$  is rewritten as  $xy$  with high probability, few other nonterminal symbols are rewritten as  $xy$ , and  $A$  appears with modest probability as a result of other rewrite rules. Negative correlations can only be seen when there is a compromise between the conflicting demands of having few occurrences of  $xy$  compared to the number of  $x$ 's and  $y$ 's, in order to make  $C_{x,y} < 1$ , and having many occurrences of  $xy$  so that the uncertainty in  $C_{x,y}$  is small enough to produce a large  $Q_{x,y}$ .

Correlations between terminal symbols do not arise only because of sequences which appear inside the rewrite rules. If  $A$  is often rewritten as  $xy$ ,  $B$  is often rewritten as  $rs$ , and  $C$  is often rewritten as  $AB$ , then  $Q_{y,r}$  will be large. Therefore one cannot blithely define nonterminal symbols using the high- $Q$  digrams in the obvious way and expect everything to work. These nonterminal symbols have to be incorporated into rewrite rules in such a way that any additional correlations which appear are desirable. Some guidance in defining rewrite rules which produce these nonterminal symbols can be obtained by calculating "higher order"  $Q$ 's such as  $Q_{xy,rs}$  and  $Q_{xy,r}$  for sequences containing high- $Q$  digrams. But much guesswork remains.

Once some of the rewrite rules are hypothesized, it may be possible to (at least partially) parse the language sample to produce a sample in which many groups of terminal symbols have been replaced by nonterminal symbols. The same sort of correlation analysis can then be applied again. The parsing may not be straightforward though. It might be hypothesized, for example, that  $A$  is rewritten as  $xy$  and  $B$  is rewritten as  $yz$ . Then it may not be clear how to parse  $xyz$ . Furthermore, reasonable probability assignments have to be guessed based on the  $C$ 's, and the  $P$ 's.

It may be useful to compute generalizations of  $C_{x,y}$  and  $Q_{x,y}$  such as  $C_{x,y,z}$  and  $Q_{x,y,z}$ , where  $C$  is defined in analogy to (6) as

$$C_{x,y,z} = P(xyz) / P(x)P(y)P(z) \quad (12)$$

and  $Q_{x,y,z}$  is again defined as  $\log(C_{x,y,z}) / \delta \log C_{x,y,z}$ . Indeed,  $C$ 's and  $Q$ 's can be defined for any partition of an  $n$ -gram. In addition to providing information which assists in guessing production rules, these general  $Q$ 's provide a way of measuring the deliberateness of a production, so that one may assess how much structure the production is providing over and above what the random grammar provides.

Even when the grammar is unknown the Q's can be useful for computing the probabilities of given strings of terminal symbols. This problem is often attacked by taking products of probabilities of substrings of the given string, in the spirit of equation (5). But where correlations exist, (5) does not hold. The Q's tell which substrings are the most and least correlated, so if the largest (in absolute value) Q's are tabulated in advance, the given string can be divided into substrings in such a way that there is little correlation between the substrings.

#### CONCLUSION

The measures of correlation which have been presented here are not billed as a grand solution to the problem of the statistical inference of grammars. But they are meant to be useful for heuristic guidelines. It may be that these measures can be used in conjunction with known heuristic search procedures to provide a useful algorithm for solving this problem. But this remains a matter for research.

1. A W Aho, and J D Ullman, The Theory of Parsing, Translation, and Compiling 1, (Prentice-Hall, 1972).
2. M R Spiegel, Theory and Problems of Probability and Statistics (Schaum's Outline Series), (McGraw-Hill, 1975), p108.
3. *ibid.* p45
4. G L Squires, Practical Physics, (McGraw-Hill, 1968), pp 34-37.