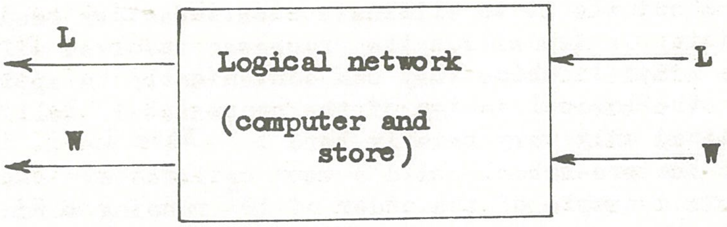
LINGUISTIC APPLICATIONS OF COMPUTING MACHINERY

by

F.W. Harwood,
Department of English, University of Tasmania
Hobart, Tasmania

This paper is concerned with the programming of computing machinery for some linguistic applications. While these concern only a part of the complex activities involved in the full use of language it is hoped that even very limited results may be of some use in connection with problems of programming the handling of information which is originally expressed in language, problems in the design of secrecy and other coding systems and problems in machine translation.

The simplest possible model we can make of the functioning of a natural language involves a logical network with language and world inputs and outputs: thus



If we consider such a network started at a certain time, for example at birth in the child language learning case, then its linguistic activity can be represented in one dimension. A suitable model is a tape divided into word positions, shifting in discrete steps, at a rate of the order of one per second from right to left. The segments of this tape may be filled in four ways:

1. by a sequence of words received;
2. by a sequence of words transmitted;
3. by a sequence of words constructed in the network but not transmitted;
4. by zero.

The main problem of the theory of language is to find the logical design of the network which generates and operates on these sequences of words and connects them with the world input and output. There is good evidence that this network operates in the parallel mode on units of more than one word in length. The nature of these units is not at present fully understood but the units of phrase and sentence structure may be taken as a starting point. Some reservations are necessary about this. On the one hand there are hierarchies of organisation concerned with larger units than the sentence. On the other hand very long sentences can usually be handled in parts. How-ever it is convenient to begin by thinking in terms of a high speed section of the network which handles the linguistic input and generates the linguistic output sentence by sentence.

The generation of a particular sentence can be regarded as a process of applying to the possible sequences of words several sets of constraints; (i) syntax, which restricts the possibilities to sequences which are used as sentences; (ii) transition probabilities; and (iii) meaning. The first two are of special importance because results in them can be derived from the structure of the linguistic signals. They thus restrict the field in which the meaning considerations have to operate. Further, results obtained from the study of them can be applied to communication problems independently of considerations of meaning; for example, to coding problems.

While the methods to be discussed are quite general they are stated in a form suited to application to sub-languages of English. By a sub-language is meant the language used in communication about a particular field of information or action. The language of pre-school children is an example of a sub-language. The aim of the restriction to a sub-language is to secure greater uniformity in the language handled. The restriction to English is made because the application of the methods involves decisions which are affected by the structure of the language considered, and because it is the only language to which the writer has applied these methods.

The first major problem is to set up in the network the syntactic system of the sub-language with which we are dealing. In the real language learning case we have a learning program which accumulates the linguistic system from the input and adjusts it to eliminate unsatisfactory results in the output. For simplicity of exposition the processes involved will be discussed separately in the order in which they can conveniently be applied. The necessary theory for the logical design of the processes is well developed and they will be outlined only very briefly here . Note 1.

In the case of the pre-school child a very satisfactory linguistic system is constructed from a sample of the order of 10^7 running words. It will be seen that to apply the processes to be discussed to samples of this order requires the use of high speed machinery. The current English language research project at the University of Tasmania consists of applying these processes to small samples by hand methods and investigating the logical design of the processes with respect to the detailed structure of English.

Let us consider the construction of a syntactic system for a sub-language of English with words w_1, w_2, ...w_r for sentences of length up to p positions. English words have a well-known frequency distribution. In the pre-school child case we may take r as around 1,000 and p as 10, and it is found that occurrences of about 300 words contribute about 90% of the text. In the case of an adult sub-language we can take r as around 5,000 with occurrences of about 1,000 words contributing at least 75% of the text. If we consider sentence length in terms of full stops in written text we cannot take p at less than 30, but we can treat most large sentences as a hierarchy of structures not involving more than 10 items on any level. As questions and commands can be obtained from statements by simple transformations we can confine our attention to the construction of a syntactic system for statement type sentences.

Let us suppose that we have a suitable source of a given sub-language. From this source we draw a sample, L', and from this sample we are to construct a syntactic system. Given r and p there are N possible 1......p positional sequences of $w_{1....r}$. We assume that there is a small sub-set, L, of these which may be used as sentences of the language, and the syntactic system is to

(1) decide whether any proposed sequence of $W_{1....r}$ is a sentence, i.e. to predict L from L', to divide the N possible sequences into L (used in the language) and L (not used in the language;)

(2) give the internal groupings of the words in the sentence, i.e. characterise the syntactic features which have a meaning residual over the word meanings. Most of these are given by the derivation of the sentence in an axiomatic syntactic system so the second requirement will

not be further discussed.

This system can be constructed as follows:*

The sentences in L' can be rearranged according to length so that we have L'_1: a sub-set of the r possible 1-positional sequences of $w_1 \ldots r$; L'_2: a sub-set of the r^2 possible 2-positional sequences of $w_1 \ldots r$, and so on to L'_p. Call this the L' array. It is an array of word patterns. The patterning is such that we can find a set of variables $c_1 \ldots m$ whose values are sub-sets of $w_1 \ldots r$ and with these, retaining certain words as constants where necessary, construct a more compact positional array called the C-array.

We can then find a set of sequences of high level variables $X_1 \ldots X_n$ whose values are sets of sequences of the C-array and store the syntactic system in the following form:

(1) A set of sequences of the variables $X_1 \ldots n$ retaining some words as constants if necessary.

(2) The systems and sub-systems which generate the C sequences which are the values of each of the variables $X_1 \ldots _n$.

(3) The word list $w_1 \ldots _r$ with indicators of which words are values of the variables $c_1 \ldots _m$.

Such a system determines a set of sequences, K, and a simple but lengthy computation yields the number of sequences in K. This will be called the sum of the system. K/N is a measure of the constraint the system applies.

The system can be expressed in its most compact form by the application of the operations of mathematical logic and there is a logical network which realises such a system.

Such a system satisfies the condition that L' is included in K. However we need to consider ways of approximating to the condition: K = L.

To do this we consider the positive and negative fit of the system to L.

Approximation to perfect positive fit, i.e. generating all the sequences in L, is given by the condition

$$L.K \quad 0 \qquad \text{or} \qquad \frac{L.K}{L} \quad 1$$

and to perfect negative fit, i.e. excluding all the sequences in L, by

$$L.K \quad 0 \qquad \text{or} \qquad \frac{L.K}{L} \quad 1$$

The variables must be inserted so that they permit not only L', but L (of which L' is but a part), and at the same time retain sufficient constraints to exclude L. The simplest approach seems to be to relax the condition of negative fit to get a system which can be used as a starting point and set in operation and adjusted.

The system may be constructed in several ways:

(i) We may design the system in the light of the available information about English syntax. A system of good positive fit can be obtained by using the usual word-classes for the variables of the C-array and following the usually recognised phrase structures in constructing the higher level variables. The sum of the system can then be computed and compared with N. (See Appendix).

* Note 2.

(2) We may experiment with processes which may be applied directly to L' to yield the system. We may begin by inserting only the variables permitted by the condition K' = L'. This will yield a large system of small classes which may be capable of some simplification. We can permit the sum to increase in arbitrary steps and substitute for sets of the small classes the larger classes in which they are included.

Study of child language learning shows that the system is first developed for short sequences. The basic sentence types are acquired and then extended by the use of longer phrase structures.

However the system is started it will be necessary to test and correct it at intervals by taking further samples of L and seeing whether they can be generated by the system, and by driving the system by random numbers and seeing whether the output is satisfactory. Corrections are made by adjusting the values of the variables, the sub-systems, or the initial set. Account may be taken of frequencies throughout the work. When these correcting processes have been more fully studied it is hoped that a program for them will be designed.

We now turn to possible applications of the method of transition tables given by Shannon*.

The term "transition tables" will be used to cover two types of table: (i) a table of transition probabilities, and (ii) a table of varieties.**

A two position table of transition probabilities gives the probability $p_i(j)$ that w_i is followed by w_j. It may also be given in the form of a diagram frequency table. A three position table gives the probability $p_{ij}(k)$ that w_i w_j is followed by w_k. Similarly to p positions. From these tables we can derive corresponding two, three ..... p position variety tables. A two-position variety, $V_i(j)$, is the set of words which occur after w_i; it is the set of words which have p o in the corresponding transition probability table. Similarly for three ..... p position varieties.

Let us consider the construction of transition tables from the sample L' discussed in connection with syntax.

1. Types of tables. Syntactic considerations suggest the following types: (a) general two, three .....n position tables from the whole sample; (b) tables for particular word positions of the form $p_i^a(\overset{b}{j})$, the probability of w_j in position b with w_i in position a, and corresponding three ... p position tables. Two cases of particular interest are (i) the tables for the successive word positions, in the two position case the tables:

$$\overset{1}{p_i}(\overset{2}{j}),\quad \overset{2}{p_i}\ (\overset{3}{j}),\ ....\overset{p-1}{p_i}\ \ (\overset{p}{j})$$

and (ii) the tables for key words at various levels of structure. We may seek a small set of key words in the sentence between which we investigate the constraints and then, from each of these, further constraints running backwards or forwards. The syntactic system provides a framework for this.

2. Size of the tables. Concern may be felt that the tables will require so much high speed storage space that they will not be a practical proposition at least at present. Variety tables can be made more compact by operations with classes. For example by forming the class of words with the same variety and by including overlapping varieties in larger classes. This is equivalent to introducing word class variables into the tables.

* Note 3 ** Note 4

3. Fit to L of a system of tables constructed from L'. We can consider the sum of the system in the same way as the sum of a syntactic system. It is obvious that the many position tables will not give sufficient syntactic freedom, in fact the p positional table is just L'. This does not satisfy the basic requirement of enabling sentences to be handled as sequences of words rather than as unstructured code groups. A system of two position tables for English does not have a perfect negative fit. But until at least two-, three-, and four-position tables have been constructed from adequate samples of text it will be difficult to go much further. Syntactic considerations suggest that these are not likely to be of unmanageable size, and they offer a way of applying a further set of constraints to the output of a syntactic system, particularly for assembling the leading words of a sentence.

These constraints may be applied in the high speed network in which the sentence is held before being transmitted.* It is interesting to note that Lashley suggests from psychological evidence that there is "a partial activation or priming of aggregates of words before the sentence is actually formulated from them ..."

Further constraints. That further constraints are involved in most uses of language is obvious. The most accessible seem to be those of logic because networks are available for some logical operations. The problem is to identify the expressions which are values of the variables of the propositional calculus, the class calculus and so on.

There are also the constraints associated with meaning which lead, for example, to the selection of a set of sentences to describe a picture, or to the obeying of a set of commands. But it may be doubted whether these constraints do more than further limit the set of acceptable responses. They are at a maximum in physical object and action situations and uses of language controlled by accepted formal rules. However the design of systems with this type of meaning response involves techniques other than those discussed in this paper.

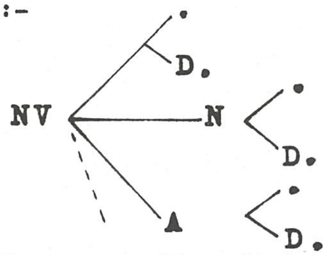
* Note 5

APPENDIX

ILLUSTRATION OF NETWORK REPRESENTING A SMALL PART OF A SYNTACTIC SYSTEM.

If we consider the English statement-type sentence using the familiar grammatical ideas of noun phrase (N), verb phrase (V), adjective phrase (A), and adverb phrase (D), etc., we find that we can generate the sentences which are used by a network of the following sort

With each verb can be associated indicators of the phrase patterns that follow it. The classification given by A.S. Hornby in "A Guide to Patterns and Usage in English" (Oxford) provides this type of information in some detail. We thus take as our initial set a tree-like structure of which the following is a small portion:-



We then provide sub-systems for generating the sequences which are values of the variables N, V, etc. For example, it is found that at many points in the structure we have a pattern illustrated by such phrases as:

Q (Quantifier)	M (Modifier)	H (Headword)
some	new	ideas
nearly all		motorists
	scientific	problems
the	very latest	technique

A compact sub-system can be designed to generate all phrases of this type. The output of this sub-system is switched to wherever it is required in the initial set, to other sub-systems generating more elaborate noun phrases in which Q M H constructions occur as parts, to a system generating adverb phrases among which are phrases of the form: preposition followed by noun phrase, and so on. Inhibitions are used to block any items not required in a particular position.

Such a system may be elaborated by taking account of probabilities at each of the points at which there are alternative paths in the network.

NOTES

1. In addition to the standard books on mathematical logic, the following papers:

   Shannon, C.E. "A Symbolic Analysis of Relay and Switching Circuits", Trans. AIEE 57.713-23 (1938)

   McCulloch, W.S., and Pitts, W. "A Logical Calculus of the Ideas Immanent in Nervous Activity", Bull. Math. Biophysics 5.115 (1943)

   Burks, A.W. and Wright, J.B. "Theory of Logical Nets", Proc. IRE (1953) 1357-65

   ed. Shannon, C.E. and McCarthy, J.: Automata Studies, Annals of Mathematical Studies No. 34 (1956).

2. Harris, Z.S. "Methods in Structural Linguistics" (1951)

   Harwood, F.W. "Axiomatic Syntax", Language 31.409 (1955)

3. Shannon, C.E. "The Mathematical Theory of Communication", B.S.T.J. (1948).

4. A method for extracting morphemes from a string of phonemes by computing varieties is given by Z.S. Harris, "From Phoneme to Morpheme", Language 31.190 (1955) .

5. Lashley, K.S.: The problem of serial order in behaviour in: "Cerebral Mechanisms in Behaviour," ed. Jeffress (1951)

DISCUSSION

Dr. J.M. Bennett, University of Sydney.

What is the meaning of the symbol "@" on your sheets?

Mr. F.W. Harwood (In Reply)

That is just to distinguish the indefinite article 'an' or 'a' from the character 'a' which is used as the conventional dictionary abbreviation for an adjective.

Dr. J.M. Bennett, University of Sydney.

I have just followed through this system and I have got a sentence out which reads 'a bad beast did not even get birds'. Does this mean that the object of your system is purely to analyse syntax and has no regard for meaning?

Mr. F.W. Harwood (In Reply)

Yes, that is the first of the constraints you apply afterwards.

Dr. S. Gill, Ferranti Ltd.

Have you made any estimate of the size of programme that would be required to carry out this syntactic analysis?

Mr. F.W. Harwood (In Reply)

To carry out the decisions required, i.e. whether a required sequence of words is a permitted sentence or not without respect to probabilities inside the system, requires only the substitution of the permitted values of the sub-system in the sub-sets.

Dr. A.S. Douglas, University of Leeds.

I would like to report some work that has been done by the Cambridge Language Translation Unit although this is not in my own field. Broadly speaking this Unit has tried to take a source language and translate it into a target language. The first thing obviously not to try and do is attempt a word-by-word translation. There appear to be, however, two approaches which can be made to the problem. One can attempt to integrate into the target language all the things that the source language has, whether the target language has them or not. For example, in French there is gender whilst in English there is not in general. For this case therefore, one would assign genders in English and hope that they were the equivalent to the ones in French. The alternative is to turn round and throw away as much as possible of the languages and try and attempt to find a base common to both. This appears to be the approach which in fact is taken by a human translator. Working on this basis therefore, the Unit attempted to throw away as much as possible and in this case the first thing that went out the window was syntax. As a result of this they set up in the machine a language which was not a language but which possessed features which were amenable to both the source and target language.

Having gone this far it seemed advisable to attempt to find a base language which was common to all languages. However, having reduced the source language into the base language it is still necessary to translate into the target language. As the machine language is basic, however, and only essentials have been retained, the formation of the translated language or the target language is a building up process and this must be done from a knowledge of the language itself and the sense of what has been presented before. This raises all sorts of very difficult problems which have as yet not been solved. The general idea of trying to build up into the target language was to use something similar to Roget's Thesaurus in which groups of words were formed into different form kinds. This idea of course is not unique and there are many types of classification which would be equally suitable. The fundamental difficulty which I think will still beat anyone in this field is that if you are trying to build up a target language from this machine language then you will need to make considerable use of context. This is a great over-simplification of the problem but it does indicate the problems that have yet to be faced in doing machine translation, and in addition to this, there is of course always the problem of uniqueness.

Dr. M.V. Wilkes, University of Cambridge.

I think Dr. Douglas that perhaps you have underestimated the advantages of the word by word translation with alternative meanings, particularly for the person who has a knowledge of the subject. After all, this is the way in which a scientist who does not know German would attempt to read a German paper. He would translate each word in the paper as it appeared, ignoring grammar, and if he knows the subject well enough he will get at the meaning. I think this should always be remembered in technological conversation and translation. In this regard the scientific user is in a favourable position compared with the professional translater. I have seen some examples of this type of thing done by Oettinger at Harvard and the results that he obtained by word listing were quite comprehensible even when translating from Russian into English, but only to those who had a knowledge of what the topic was supposed to be. I would suggest therefore that, since this is of value to technologists and scientists, this is perhaps the approach that should be made initially in language translation. The other requirements would then follow as a result of experience with this system.

Mr. F.W. Harwood (In Reply)

I agree that these comments are quite justified but after all you must investigate what are the methods and systems which can be used to try and provide a literal, elegant, translation.