XV. MECHANICAL TRANSLATION

Dr. V. H. Yngve

A.. A NEW PUBLICATION

In cooperation with Prof. W. N. Locke of the Department of Modern Languages, M.I.T., a new publication called MECHANICAL TRANSLATION is being published on a trial basis for one year. During the year, issues will appear at irregular intervals as material becomes available. The purpose of the publication is to provide this new and rapidly growing field with a much needed means of communication and exchange of ideas. The first issue, which appeared in March, contains an annotated bibliography of 41 entries.

B. LANGUAGE AS AN ERROR CORRECTING CODE

Workers in the field of mechanical translation have expressed surprise that a word-for-word translation is as good as it is. Their intuitive feeling apparently was that it would be next to worthless. Experiments in which the words of the foreign language have been replaced by lists of all possible English equivalents for each word would give millions or billions of possible sentences based on the various choices of meanings; yet a person can find his way through such a maze with fair facility and pick out the particular meanings that are best. Such experiments have focused attention on the error correcting nature of language.

The error correcting codes that have been investigated in information theory literature operate by recoding the original message in a special way involving constraints between successive symbols so that an error in one or more of the received symbols can be corrected and the original message restored. One way in which this can be understood is to consider all possible sequences of symbols as possible received sequences. Now if only a fraction of these possible sequences is permitted to be transmitted as messages, we have introduced a form of constraint. If the permitted sequences are different enough, an error in one or more symbols of one of these sequences will not transform it into another permitted sequence. In fact, it may still be different enough from all other permitted sequences to be identifiable by an appropriate recovery technique. The constraints that are effective in providing the error correcting nature of the code have nothing to do with the relative transmission probabilities of the various permitted message sequences. The important thing is that some sequences of symbols are allowed and some are not.

It is interesting, in the light of the demonstrated error correcting properties of language, to investigate the amount of the redundancy of language that is due to the fact that certain sequences are virtually impossible, and the amount that is due to the fact that the allowed sequences have widely different probabilities. Consideration of the number of letter sequences actually assigned to words in any given language reveals that most

of the redundancy at the word level is due to the fact that only a very small fraction of the possible letter sequences are actual words. Only a small amount of the redundancy is contributed by the widely varying probabilities of the allowed words. For example, Shannon calculates that the entropy is 11.82 bits per word on the basis of a vocabulary of 8727 words and a law of word probability given by $p_n = 0.1/n$, where n is the rank of the word when ordered by frequency. If these 8727 words were considered equiprobable, the entropy would be $\log_2 8727 = 13.08$ bits per word. The small difference, 1.26 bits per word, can be attributed to the effect of the hyperbolic law of probability used.

There is, however, an outstanding difference between the way in which constraints. are incorporated in language and the way in which they have been incorporated in error correcting codes in the past. Constraints are incorporated in language on various levels. For example, out of all possible letter sequences, very few are actually assigned to words. Out of all possible word sequences, only a few are actually assigned to phrases. Cut of all possible phrase sequences, only a few are actually assigned to sentences. Out of all possible sentence sequences, only a few make coherent paragraphs. The fact that language is structured on different levels aids in reconstructing the original of a garbled message. At the word level, one can frequently decide whether a word has been misspelled, or whether there has been a misprint. At the phrase level, one can tell if a word has been transformed into another meaningful word, and so on. This property of incorporating the constraints at various levels makes it possible to get a certain amount of error correction immediately at the lowest levels, and then to have additional error correcting ability as the message grows longer. Occasionally we do not decide what a given word must have been until several sentences have gone by. The efficiency of these language constraints in providing error correction has not been thoroughly investigated.

On the other hand, error correcting codes in the past have worked on only one level. The material to be transmitted has been divided into blocks of a given length and recoded into a code with special constraints which can correct a certain number of errors in each block. However, errors are usually produced at random, and the number of errors per block is not constant, but follows a Poisson distribution. Thus with an average of one error per block, a one-, three-, or five-error correcting code will leave uncorrected over 26 percent, nearly 2 percent, or over 0.05 percent of the blocks, respectively. It has proved difficult to construct high-order error correcting codes with very long blocks, and they would prove unwieldy in practice.

It appears that error correcting codes could be designed with advantage on a layer principle, as is language. One could correct single errors on the lowest level. On the next level, the multiple errors could be corrected on a word basis with fewer additional symbols needed than on the lowest level. The process could be continued in this manner up through the layers. In section IX, Prof. P. Elias has worked out such a code and has shown it to possess certain definite advantages.