Ann. Bull. RILP, No. 12, 127-140 (1978)

# TRANSMISSION OF MEANING BY LANGUAGE

Hiroya Fujisaki, Keikichi Hirose* and Yasuhiro Katagiri*

## Abstract

As a step toward quantifying meaning and the process of its transmission by language, an experiment was designed to facilitate observation of the process of expression by the sender as well as the process of comprehension by the receiver. By confining the subject matter to ages and restricting the available vacabulary to a small set of nouns, characteristics of the sender and the receiver were measured and formulated quantitatively. A model was then constructed for the entire process through which meaning is transmitted by a word, clarifying various causes of ambiguity and inaccuracy of transmission. The whole process was evaluated in terms of the amount of information transmitted and the r. m. s. error of transmission.

## 1. Introduction

Since language is the most important medium of thought and communication, the study of language has a long history. The majority of the classical studies on language has been oriented, however, toward analysis and description of the structure of a language *per se* as a system of codes, or toward comparison of structures of various languages. Hence little has been studied on language use, i. e., the relationship between language as a code and the information to be transmitted and received through the use of language. This situation may be ascribed primarily to the fact that problems of language use have been traditionally considered to belong to the realm of psychology rather than to that of linguistics. The tendency still persists at present, and the studies of semantics and pragmatics are scarcely conducted with an awareness of the underlying psychological processes. On the other hand, a major part of present research efforts in psycholinguistics is focused on the psychological procedures or perceptual strategies adopted in parsing syntactic structures, and comparatively little is accomplished by way of understanding the processes of coding and decoding, i. e., expression and comprehension of the meaning of linguistic messages.

The process of communication by language may quite generally be represented by Fig. 1. In this process, a certain part of the information $I_1$, which one person (the sender) possesses and intends to transmit, is transformed into a linguistic expression E and is presented to another person (the receiver). Upon receiving the expression E, the receiver acquires the information $I_2$ which generally is an approximation to the information $I_1$ intended by the sender. In ordinary situations, the linguistic expression E is always converted into a physical signal for the purpose of

* Department of Electrical Engineering, Faculty of Engineering, University of Tokyo.

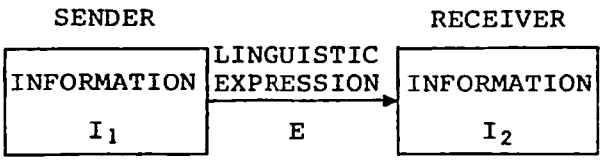


Fig. 1. The process of communication by language.

transmission, and lends itself to quantitative measurement and objective description. The information contents $I_1$ and $I_2$, on the other hand, generally elude direct observation, since they are only represented by the psychological states of the sender and the receiver. Quantitative descriptions of the sender's coding characteristics and the receiver's decoding characteristics are not possible under these circumstances.

In order to bring these coding and decoding processes into more tractable forms, we shall adopt a situation where an object $O_1$ with a physically observable attribute is presented to a subject, who as a sender selects an expression E to describe the object $O_1$. The expression E is presented to another subject, who as a receiver selects, among various alternatives, an object $O_2$ which he considers to be implied by the expression E. The situation is schematically shown by Fig. 2. The addition of physically

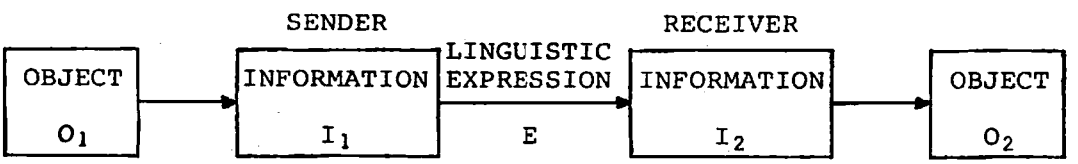


Fig. 2. The process of communication by language where physically observable objects are added to facilitate quantitative descriptions of sender and receiver characteristics.

observable objects at both ends of the communication process thus makes it possible to measure the coding and decoding processes by the method of experimental psychology, and to describe their characteristics in quantitative terms, though it imposes certain limitations on the nature of information to be transmitted.

The present paper describes an approach to the quantitative formulation of the process of transmission of meaning based on the above considerations. Experiments are designed to measure both the coding and the decoding characteristics of a subject in situations where a word is used to transmit information. Based on these measurements, a quantitative model is presented for the process of semantic information transmission, and a method is shown for the evaluation of the ambiguity and the accuracy of transmission.

## 2. The Model

As already stated in the previous section, the present investigation will be restricted to situations where a physically measurable attribute of objects, such as the age of a person or the color of a paint, constitutes the information to be transmitted through the use of language. The linguistic expressions to be used for transmission will also be restricted to a predetermined set of nouns. The psychological processes involved in such communication situations are shown in Fig. 3. When the physical attribute is presented as a stimulus to the sender, it is converted into a percept on a certain perceptual continuum. The selection of a linguistic expression is based on quantization and coding of the perceptual continuum. Thus the sender's process of verbalization can be considered as consisting of two sub-processes: perception of a stimulus and its expression by language. Since the intervening percept eludes direct measurement, however, these two sub-processes shall be treated as a single process and its input-output characteristics shall be defined as the sender's coding characteristics throughout the rest of this paper.

The linguistic expression (a noun in this case) selected by the sender is transformed into letter strings, and is presented as a stimulus to the receiver. The perceived expression is decoded by the receiver and reconstructs a percept in the receiver's mind, which is an approximation to the one in the sender's mind. While information transmission is accomplished and communication in the ordinary sense is terminated by this decoding process, another process has to be added in order for the receiver's percept to be transformed into a measurable entity. Namely, the receiver is asked to reproduce the original stimulus which is implied by the received expression. The reproduction is accomplished by selecting a physical stimulus from a number of candidates. Thus the receiver's process of understanding a message consists of two sub-processes: comprehension and reproduction. As in the case of the sender's coding characteristics, these two sub-processes shall be treated as one and its input-output characteristics shall be defined as the receiver's decoding characteristics.

These coding and decoding processes are in some respects decision processes. Since human decisions are not exempt from statistical fluctuations caused by a number of psychological and physiological factors, characteristics of these processes may most properly be described in probabilistic terms. In other words, a sender's coding characteristics can be represented by the set of probability distributions that each one of the possible expressions displays on the continuum of stimuli presented to the sender. On the other hand, a receiver's decoding characteristics can be represented by the set of probability density functions of the stimuli reproduced by the receiver against each one of the transmitted expressions.

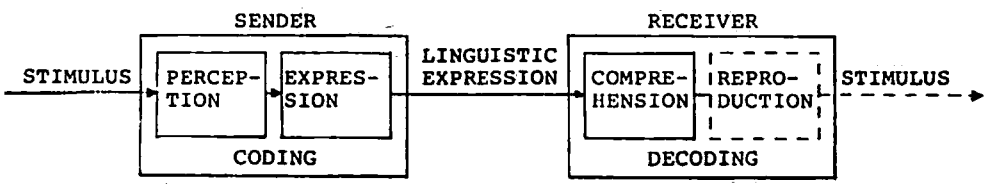


Fig. 3. Processes in transmission of meaning by language in the communication situation of Fig. 2.

## 3. Experiment

### 3.1 Method

As an example of experimental determination of the coding and decoding characteristics involved in the above-mentioned model of semantic information transmission, an investigation was conducted on the relationship between the age of a person and the nouns of Japanese commonly used to designate the age. Most languages, including Japanese, provide nouns for classifying and designating age ranges, but the age ranges they designate are not necessarily equal in size, nor complementary in their distribution. Moreover, the information conveyed by these nouns is not restricted to the chronological age of a person, and the use of these nouns is influenced by a number of contextual factors. In order to minimize the effects of these extraneous factors, a set of nouns were selected such that they were nearly complementary with each other in designating age and nearly uniform in other aspects, and the measurement of coding and decoding characteristics were conducted using this pre-determined vocabulary. The vocabulary used in the major part of the following experiments consisted of the five Japanese nouns: "yō-nen", "shō-nen", "sei-nen", "sō-nen", and "rō-nen", corresponding roughly (but not exactly) to the English "childhood", "boyhood", "youth", "manhood", and "old age", respectively. In some part of the experiments, however, the size of the vocabulary was controlled to see the effect.

#### Subjects

A total of nine subjects, eight male adults and one female adult, took part in the following experiments. They were all native speakers of Japanese and their ages ranged from 22 to 47. Each subject served both as a sender and a receiver. Since individual differences are naturally found in characteristics of language users, the experimental data of the nine subjects were not pooled, but were analyzed individually to extract coding and decoding characteristics of each subject.

#### Measurement of coding characteristics

The measurement of a sender's coding characteristics was conducted by the method of constant stimuli. A randomized list of integers from 0 to 70 was presented sequentially to a subject by a digital computer. The subject was instructed to select, by forced judgment, a noun from a pre-determined vocabulary which he or she considered to be appropriate for the age represented by the integer. Each subject made at least 10 responses to each integer presented, and the individual data were analyzed to obtain the probability distributions of the words on the age scale.

#### Measurement of decoding characteristics

The measurement of a receiver's decoding characteristics was conducted also by the method of constant stimuli. A randomized list of nouns in a pre-determined vocabulary was presented to a subject by a digital computer. The subject was instructed to select, by forced judgment, an integer which he or she considered to be appropriate for the noun presented.

Each subject made at least 100 judgments on each of the nouns, and the individual data were analyzed to obtain the probability density functions of the ages reproduced from the nouns.

The measurements of coding and decoding characteristics can naturally be conducted independently, but the results can be combined to study the communication process where information is transmitted between an arbitrary pair of subjects.

## 3.2 Results

### Coding characteristics

As an example of a sender's coding characteristics, the performance of one subject in the five-category coding experiment is shown in Fig. 4.

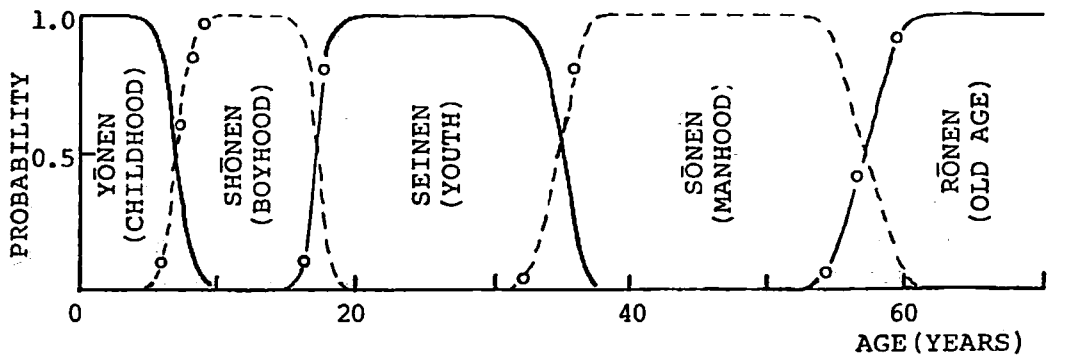


Fig. 4. An example of a sender's coding characteristics.

It can be seen that the probability distributions of the five nouns are contiguous, i.e., that the decay of the probability of occurrence of one category is accompanied only by the rise of the probability of another category, and no more than two categories occur at any point on the age axis. In this case, the categorical judgment, i.e., the choice of a particular category against a given stimulus, can be regarded as a binary threshold operation whose threshold fluctuates due to a number of psychological and physiological factors. Hence the probability distribution of a category may be approximated, in the vicinity of the category boundary, by a Gaussian distribution. The validity of the approximation is demonstrated by Fig. 5, where the probability of occurrence of "rō-nen" for the same subject is plotted on the normal scale against the age. Thus a transition from one category to another can be characterized by the mean $\theta$ and the standard deviation $\rho$, to be defined respectively as the coding boundary and the coding accuracy, of the probability distribution for one of the categories. The coding characteristics of a sender can then be characterized by the set of $\theta$'s and $\rho$'s for all the category boundaries. Quite naturally, individual differences are to be expected in the values of these parameters.

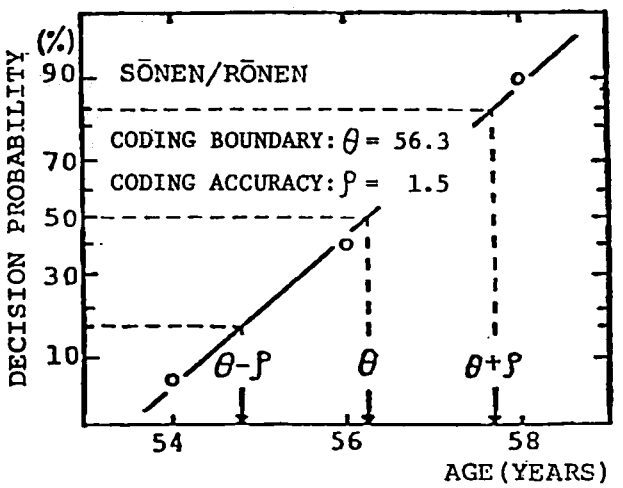


Fig. 5. Approximation of coding characteristics by a Gaussian distribution.

Individual differences in coding boundaries are illustrated by Fig. 6, where the standard deviation S ( θ ) of each of the four coding boundaries is calculated for the nine subjects, and is plotted against the mean m ( θ ).

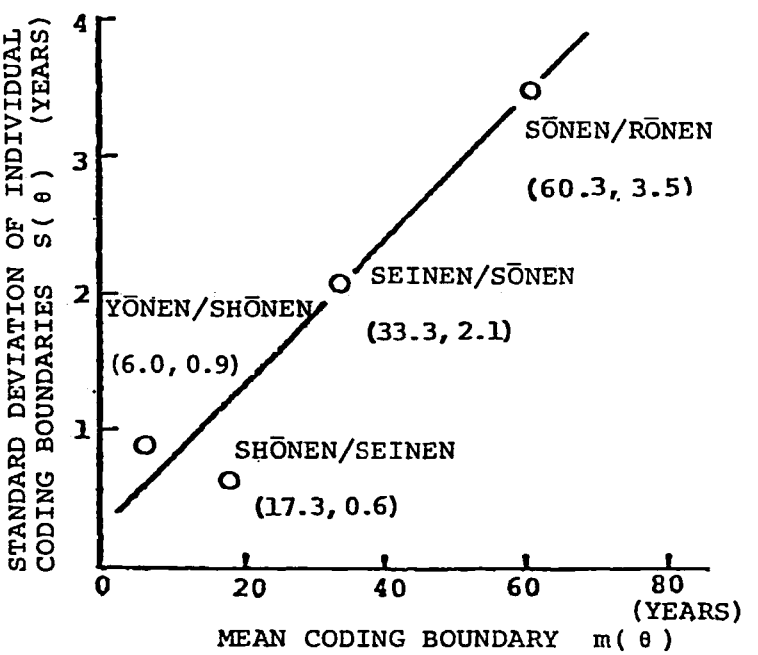


Fig. 6. Individual differences in coding boundaries in nine subjects.

Individual differences in the coding boundary, as represented by S ( θ ), are generally seen to increase almost in proportion to m ( θ ), but are exceptionally small for the boundary between "shō-nen" and "sei-nen". Fig. 7 shows the mean coding accuracy m ( ρ ) at each of the four coding boundaries, calculated for the nine subjects and plotted against the same abscissa as in Fig. 6. While m ( ρ ) is less than one year at the first two boundaries (i. e., yō-nen/shō-nen and shō-nen/sei-nen), it is approximately two years at the other two boundaries (i. e., sei-nen/sō-nen and sō-nen/rō-nen).

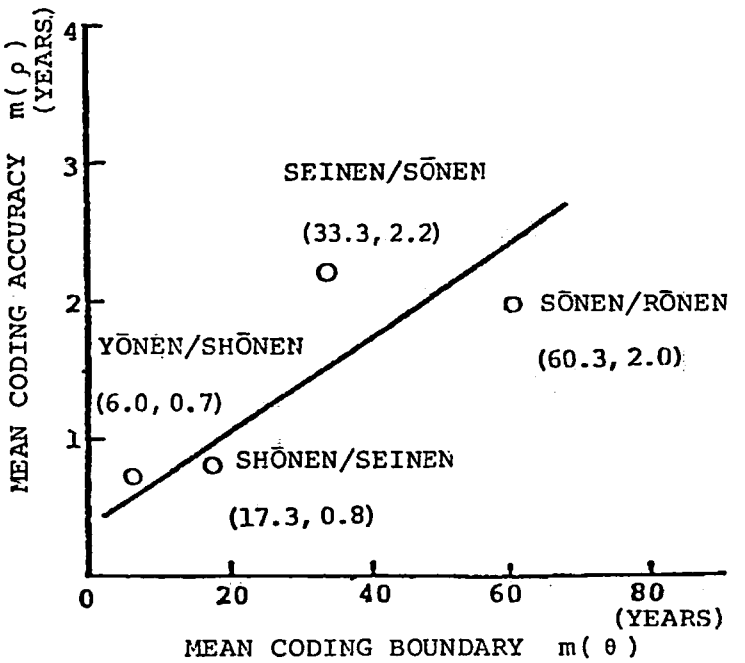


Fig. 7. Mean coding accuracy versus mean coding boundary in nine subjects.

Decoding characteristics

As an example of a receiver's decoding characteristics, the performance of one subject in the five-category decoding experiment is shown in Fig. 8. The decoding characteristics are represented by the set of five

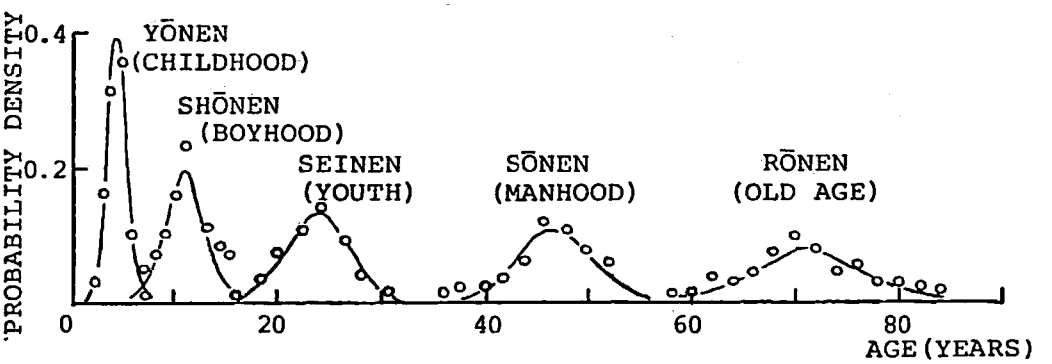


Fig. 8. An example of a receiver's decoding characteristics.

probability density functions corresponding to the five categories, and are quite different from the coding characteristics. If we assume that a receiver's response to a noun is based on a certain psychological reference but is perturbed by a number of psychological and physiological factors, the probability density function may be approximated by a Gaussian density function. The validity of the approximation is demonstrated by Fig. 9, where the cumulative probability for the response to "sei-nen" is calculated from the decoding data of the same subject and is plotted on the normal scale against the age. Thus a receiver's response to a particular noun can

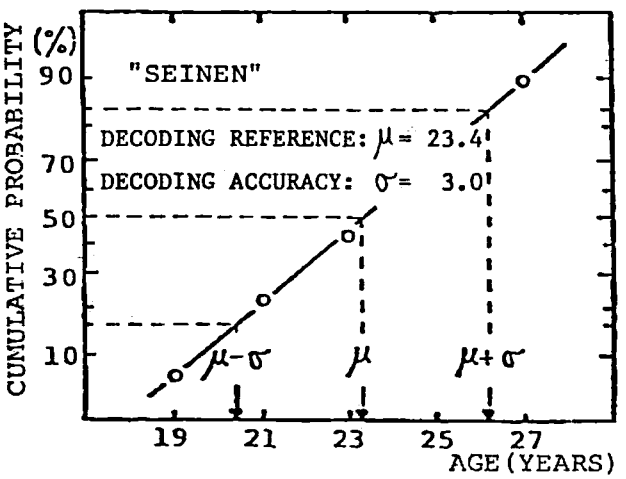


Fig. 9. Approximation of decoding characteristics by a Gaussian distribution.

be characterized by the mean $\mu$ and the standard deviation $\sigma$, to be defined respectively as the decoding reference and the decoding accuracy, of the probability density function. The decoding characteristics of a receiver can then be characterized by the set of $\mu$'s and $\sigma$'s for all the noun categories. As in the case of coding characteristics, individual differences are to be expected in the values of these parameters.

Individual differences in decoding references are illustrated by Fig. 10, where the standard deviation S ( $\mu$ ) of each of the five decoding references is calculated for the nine subjects, and is plotted against the mean m ( $\mu$ ). The standard deviation S ( $\mu$ ), as an index for individual differences in $\mu$, is seen to be less than one year for the first two categories, but is nearly equal to two years for the other three categories. Figure 11 shows the mean decoding accuracy m ( $\sigma$) for each of the five categories, calculated for the nine subjects and plotted against the same abscissa as in Fig. 10. A monotone relationship is observed to exist between the mean decoding reference m ( $\mu$ ) and the mean decoding accuracy m ($\sigma$).

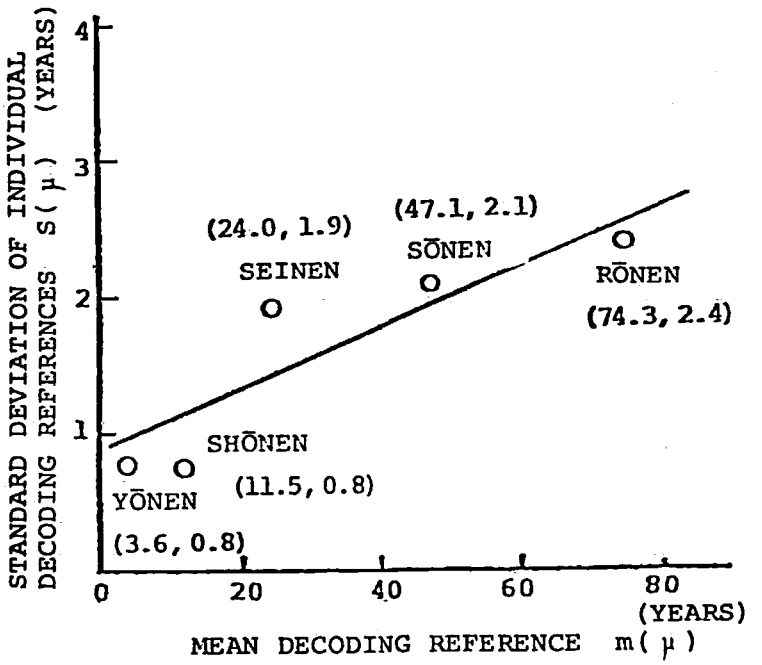


Fig. 10. Individual differences in decoding references in nine subjects.

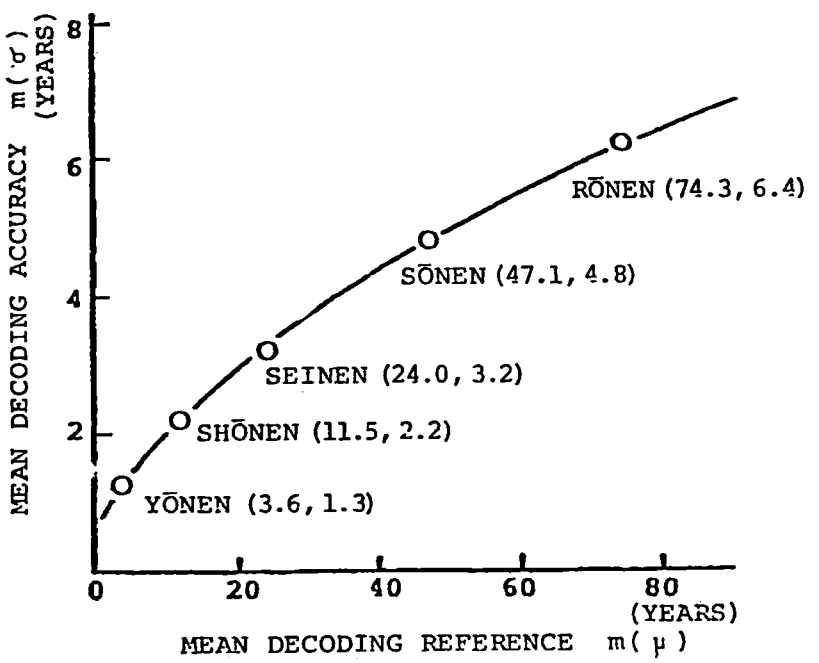


Fig. 11. Mean decoding accuracy versus mean decoding reference in nine subjects.

Discrepancy between coding and decoding characteristics

As shown by the above examples, the coding and decoding characteristics are apparently different. The coding characteristics of Fig. 4 indicate that a sender assigns a noun to any stimuli within a fairly wide range with almost equal probability. On the other hand, the decoding characteristics of Fig. 8 indicate that the stimuli reproduced by a receiver from a noun tend to be concentrated around a certain reference within the range. Essentially, a sender's coding process may thus be regarded as a quantization process, and a receiver's decoding process may be regarded as a clustering process, both in the broadest sense of the words. It may easily be observed that the stimulus $S_2$ reproduced by a receiver from an expression E is generally different from the stimulus $S_1$ which a sender tries to express by E. For the purpose of the present study, we may call the stimulus $S_1$ "the meaning of E implied by the sender," and the stimulus $S_2$ "the meaning of E recovered by the receiver."

While a discrepancy between $S_1$ and $S_2$ cannot generally be avoided, it is helpful to know the mean discrepancy to be expected in a number of cases. As the index for the averaged meaning implied by a sender, we adopt the center of each coding step, i.e., the mean value of two adjacent coding boundaries. The uppermost coding boundary for the "rō-nen" is arbitrarily set at 85. The decoding reference $\mu$ is adopted as the index for the averaged meaning recovered by a receiver. The relationship between these two indices is shown in Fig. 12, where the symbol "+" indicates the mean of nine subjects, and a rectangle indicates the region within which performances of all the possible sender-receiver combinations from the nine subjects are expected to occur. The mean discrepancy between the coding and decoding characteristics, averaged over all nine subjects, is found to be quite small.

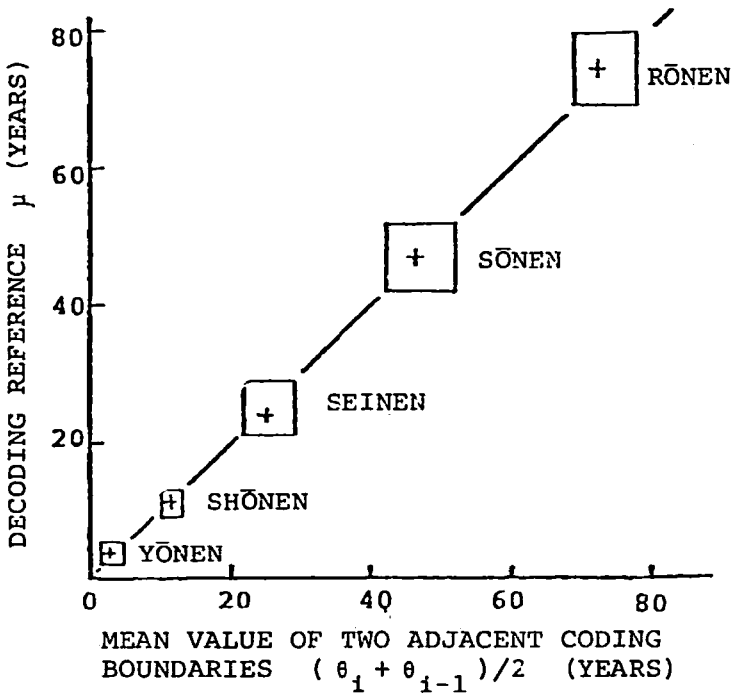


Fig. 12. Discrepancy between sender and receiver characteristics.

Effects of modifiers upon decoding characteristics

In order to obtain neutral and unbiased characteristics of coding and decoding, and thus to find out the "meaning" of a word itself, the foregoing experiments were conducted by completely eliminating the context. The meaning of a word, however, certainly depends on the context in which it is used. As an attempt to quantify such contextual effects, the following experiment was conducted to measure the influence of an adjective upon the decoding characteristics. Except for the addition of an adjective before the noun, the method was the same as that adopted in the foregoing experiments, and the number of subjects was five. Here only the noun "sei-nen" (youth) was used, and the adjective was either "ubuna" (naïve) or "rō-sei-shita" (mature). Figure 13 shows the results of these two cases as probability density functions, together with the one obtained without an adjective.

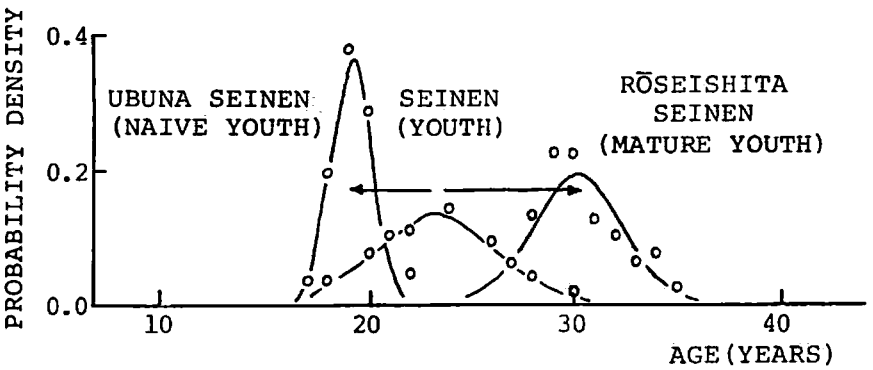


Fig. 13. Influence of modifiers on decoding characteristics.

The two adjectives are probably the ones which exert the strongest influence upon the meaning of the word "sei-nen", and preliminary experiments using other adjectives show some intermediate results. These results indicate that the effect of modifiers can be quantitatively represented by the shift of the decoding reference $\mu$ and the change in the decoding accuracy $\sigma$.

4. Formulation of the Process of Semantic Information Transmission

The process of semantic information transmission by means of a word can be represented schematically by Fig. 14. The coding process at the sender can be considered as consisting of probabilistic quantization and subsequent assignment of codes. The quantization is probabilistic in the sense that each threshold $\theta_i$ is perturbed by a random noise $m_i$. On the other hand, the decoding process at the receiver can be considered as probabilistic reproduction of the quantization level corresponding to the received code, with the addition of a bias $\beta_i$ and a random noise $n_i$. The bias $\beta_i$ here represents the difference between the center of the i-th coding step $(\theta_i+\theta_{i-1})/2$ of the sender and the i-th decoding reference $\mu_i$ of the receiver. The reproduction is also probabilistic in the sense that each decoding reference is perturbed by a random noise $n_i$. The three variables x, w, and y in Fig. 14 respectively represent the meaning implied by the sender, the code word adopted for transmission, and the meaning recovered

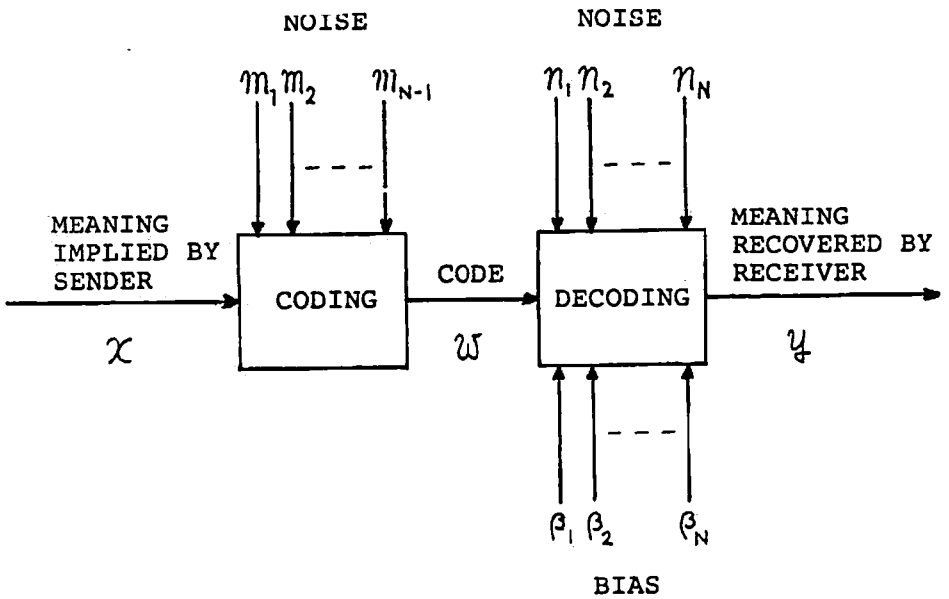


Fig. 14. Formulation of the process of semantic information transmission.

by the receiver. The entire process as a communication channel can be characterized by the conditional probability $q(y|x)$.

Based on these formulations, it is possible to evaluate the mutual information R (in bit/word) and the r. m. s. transmission error E (in units of age) for a given probability density $p(x)$ of the stimulus x presented to the sender. They are given respectively by

$$R(p,q)=\int\int p(x)\,q(y|x)\log[q(y|x)/q(y)]\,dx\,dy, \tag{1}$$

where

$$q(y)=\int p(x)\,q(y|x)\,dx,$$

and

$$E(p,q)=\left[\int\int p(x)\,q(y|x)\,(y-x)^2\,dx\,dy\right]^{1/2}. \tag{2}$$

The mutual information R may be regarded as an index for the ambiguity, and the r. m. s. transmission error E may be regarded as an index for the accuracy of transmission. It is to be noted that a larger value of R indicates a smaller ambiguity, while a larger value of E indicates a lower accuracy of transmission.

These quantities were calculated from the measured data of coding and decoding characteristics of three subjects, obtained for the vocabulary size of two, five and eight. The calculation was made only on those cases where the same subject was assumed to be both the sender and the receiver, and the probability distribution $p(x)$ of the stimuli was assumed to be uniform. These assumptions were made only for the sake of simplicity. The results

are shown in Fig. 15, where a small circle indicates the mean of three cases and a vertical line segment indicates their range of distribution. The results indicate that improvements in ambiguity and accuracy of transmission are possible by increasing the vocabulary size, but at the same time suggest that there is a limit to the accuracy of transmission attainable by the increase of vocabulary size. Thus there seems to exist an optimum size of vocabulary from the point of view of transmission efficiency.

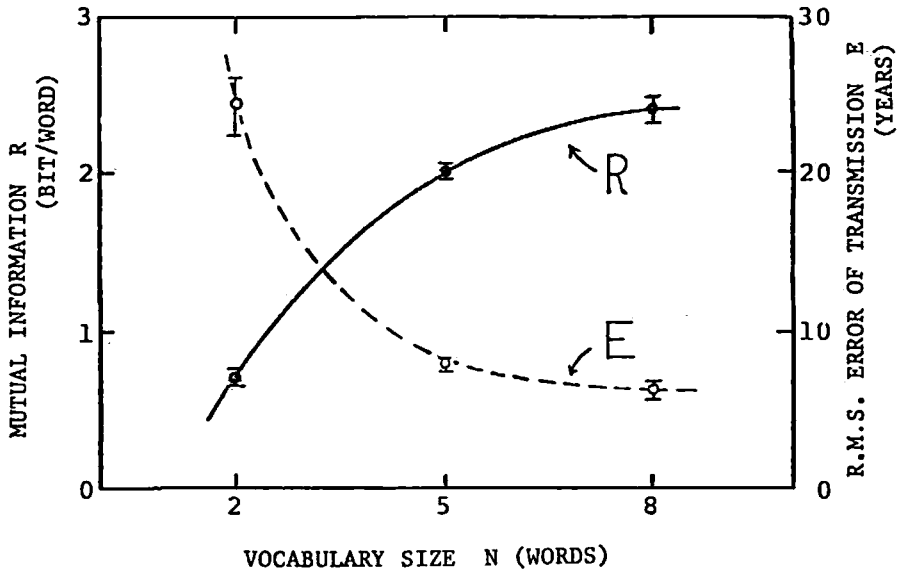


Fig. 15. Ambiguity and accuracy of transmission versus vocabulary size.

5. Comments

To the best of the present authors' knowledge, the studies by Lenneberg and his co-workers are almost the only ones which have treated the problem of language use from a probabilistic point of view. It may not be an overstatement to say that these studies represent a pioneering effort to introduce methods of experimental psychology into the study of the transmission of meaning. The use of unrestricted vocabulary and the pooling of data by individual subjects, however, did not allow these authors to attain a clear insight into the underlying process of the coding behavior of a language user. It should also be mentioned that their methods to associate meaning with words failed to reveal the essential characteristics of the decoding process in a receiver.

The formulation of a functional relationship between a physical variable and a linguistic expression is treated also by Zadeh, though quite qualitatively, as an example in his proposal for the concept of "fuzziness." The membership function is defined, not as the probability, but as the "grade of membership" that a stimulus is considered to belong to a "fuzzy" set. The present authors consider, however, that the membership function proposed by Zadeh is exactly equivalent to the probability distributions in the coding characteristics obtained by observing the sender's behavior. Despite the conceptual distinction made by Zadeh between probability and

membership function, the procedure of determining the membership function might be exactly the same as that of determining the coding characteristics described in the present study. It should also be noted that the concept of fuzziness corresponds only to one kind of indeterminacy in language use, i.e., the indeterminacy in regards to coding. There exists another kind of indeterminacy, i.e., indeterminacy in regards to decoding. The formulation of indeterminacy in language use cannot be complete unless both of these factors are taken into account.

## 6. Conclusions

For the purpose of quantifying the transmission of meaning by language, experimental techniques were developed to observe both a sender's process of language expression and a receiver's process of language comprehension. The characteristics of these processes were measured and formulated. Based on these results, a model was then constructed for the process of semantic information transmission by means of a word, and a method is shown to evaluate the process in terms of the ambiguity and the accuracy of transmission.

Although the results shown in this paper are quite limited, the techniques proposed here are applicable to a wide range of problems of language use, such as the quantitative analysis of individual, contextual, dialectal and language differences in semantic information transmission.

## Acknowledgment

The research reported here was supported by a Grant-in-Aid for Scientific Research (No. 310708) from the Ministry of Education.

## References

1. Brown, R. W. and E. H. Lenneberg (1954); A Study in Language and Cognition, J. Abnormal and Social Psychology, 49, 454-462.
2. Fujisaki, H., K. Hirose, Y. Katagiri, and W. Takeuchi (1978); Quantification of Processes in Transmission of Meaning by Language, Report of Technical Group on Automata and Languages, IECE of Japan, AL 77-77.
3. Lantz, D. and E. H. Lenneberg (1966); Verbal Communication and Color Memory in the Deaf and Hearing, Child Development, 37, 765-779.
4. Lenneberg, E. H. (1967); Biological Foundations of Language, John Wiley & Sons.
5. Osgood, C. E., G. J. Suci, and P. H. Tannenbaum (1957); The Measurement of Meaning, Univ. of Illinois Press.
6. Shannon, C. E. and W. Weaver (1949); The Mathematical Theory of Communication, Univ. of Illinois Press.
7. Zadeh, L. A. (1965); Fuzzy Sets, Information and Control, 8, 338-353.