Journal of Combinatorial Mathematics
and Combinatorial Computing
www.combinatorialpress.com/jcmcc

Structured Thinking and Information Transfer in Foreign Language Speech and Interpretation

Ling Gu[1,✉]

[1] *School of Humanities and Law, Gannan University of Science and Technology, Ganzhou, Jiangxi, 341000, China*

ABSTRACT

In this paper, information theory and information metrics are used to obtain an approximate estimation of linguistic information entropy. After that, the binary model of large-scale corpus and foreign language words is established, N-Gram model is constructed, and the information entropy of modern foreign language speech is estimated. Finally, the N-Gram model was utilized to statistically analyze the results of interpreting information loss, comparing the rate of information transfer in foreign language speeches and the subjects' interpreting performance. The results showed that the phenomenon of information loss was prominent, with many types of loss, high frequency, and serious loss situations. T assertions had 8.61%-18.95% of propositional information loss, 3.0%-7.6% of constituent information loss, and 49.68% of overall loss. The data on the information loss of each language component showed that TPO and SPE presented the most and the least frequency among the 6 propositional information losses, which were 67 and 1 times, respectively. Among the 13 types of information component loss, TFLS presented the highest frequency and TLE and SFLO presented the lowest, with their losses of 55, 1, and 1 times, respectively. In the interpreted material of English speech, the rate of narration was 2.25 words per second and the average rate was 13.45 bits per second. Among the T assertions, numbers S7, S4, and S9 have the highest propositional untranslated rate (21.8%), propositional mistranslated rate (23.5%), and propositional information loss rate (44.5%), respectively; the corresponding lowest values are at S4 (2.7%), S5 (1.8%), and S4 (2.8%).

Keywords: Information Theory, Information Entropy, N-Gram Model, Foreign Language Speech and Interpretation

1. Introduction

With the increasingly frequent international exchanges and trade transactions, the status of English as an international lingua franca is becoming more and more important. English speech is a discipline that is highly valued by western societies, and many European and American students can make fluent

✉ Corresponding author.
E-mail address: alexgu2005@163.com (L. Gu).
Received 03 March 2024; Revised 20 April 2024; Accepted 10 November 2024; Published Online 15 April 2025.
DOI: 10.61091/jcmcc127a-213

English speeches with very clear logic in middle school [1-4]. English speech is a kind of social activity which is expressed through the language of English. Because of the need to examine the use of language and to train the body language during the speech, English speech has high requirements on the speaker's language literacy and body language, thus making it an art [5-8].

Interpretation and presentation skills in the verbal and non-verbal levels have great similarities and differences, interpreting ability and presentation skills are inseparable. Interpretation is divided into formal and informal interpreting. The interpreter should be an excellent speaker first, otherwise, even if the translation is excellent, the audience will be bored because of poor sound quality, rhythm or diction [9-12]. Obviously, for a successful interpreter, mastering solid presentation skills is fundamental. Especially in the formal environment, interpreting is carried out under the public's gaze, which requires the interpreter's language expression, mental ability and body language ability to be more strict [13-16]. This makes the requirements of formal interpreters and English speakers on their respective qualities a little bit similar. Therefore, an excellent interpreter is not only proficient in both languages, but also has good speaking ability, which is crucial for interpreting work [17-20].

This paper first introduces the role of information theory, information entropy and information metrics in the identification of information in the process of structured thought interpretation, and then realizes the approximate estimation of linguistic information entropy. Then, using the N-Gram model to first establish a large-scale foreign language corpus, the information entropy of modern foreign languages estimated by different language models is obtained, so as to provide a deeper understanding of information theory metrics and data mining. Finally, the loss generated in the interpretation process, the rate of information transfer and the subjects' interpretation performance are investigated and analyzed.

2. Information transfer and application of structured thinking in interpretation

2.1. Information Theory and Information Metrics

2.1.1. Information theory and information entropy. Information theory [21] is an important fundamental theory for the study of communication transmission, signal processing, information extraction and utilization. Information theory is closely related to a range of applied disciplines and is widely used in areas such as communication coding, speech and image signal compression, computer file compression, digital communication systems, pattern recognition and machine learning.

Assuming that random variable $X=\{x_1,x_2,...,x_m\}$ has a probability distribution of $P(X)=\{P(x_1),P(x_2),...,P(x_m)\}$, the information entropy of random variable x can be expressed as:

$$H(X)=\sum_{i=1}^{m}P(x_i)\log\frac{1}{P(x_i)}=-\sum_{i=1}^{m}P(x_i)\log(P(x_i)) \tag{1}$$

Where, $P(x_i)$ denotes, the probability that X takes the value of x_i. Information entropy [22] has many important properties as a measure of information. The information entropy is maximum when the random variable X has equal probability of taking each value, when the uncertainty of the random variable X is at its maximum.

2.1.2. Information entropy and discriminating information. The discriminative information, also known as KL scatter [23], is a generalized form of information entropy. Suppose a random variable $X=\{x_1,x_2,...,x_m\}$, which has two probability distributions $P_1(X)=\{P_1(x_1),P_1(x_2),...,P_1(x_m)\}$ and $P_2(X)=\{P_2(x_1),P_2(x_2),...,P_2(x_m)\}$ under two different assumptions H_1 and H_2. Then the discriminative information or KL scatter about random variable X under two possible probability distributions $P_1(X)$ and $P_2(X)$ is defined as:

$$KL(P_2, P_1; X) = \sum_{i=1}^{m} P_2(x_i) \log \frac{P_2(x_i)}{P_1(x_i)} \tag{2}$$

where $KL(P_2, P_1; X)$ can be abbreviated as $KL(P_2 \| P_1)$, denotes the amount of information obtained by averaging the observations of random variables X under the distribution of hypothesis H_2 in favor of H_2 for the purpose of identifying H_2 and H_1. It can also be interpreted as the amount of information obtained when the knowledge of the random variable X is distributed by the probability distribution $P_1(X) \to P_2(X)$. In this case, $P_1(X)$ can be viewed as the prior probability distribution for random variable X, and $P_2(X)$ can be viewed as the posterior probability distribution obtained for random variable X. KL scatter can be a measure of the "distance" between two probability distributions for the same random variable to some extent. However, the KL scatter is directional, in general $KL(P_2 \| P_1) \neq KL(P_1 \| P_2)$.

Assume that the actual probability distribution of random variable X is $P(X) = \{P(x_1), P(x_2), ..., P(x_m)\}$ and that the information entropy with respect to X is $H(X)$. Consider an equal probability uniform distribution $P_u(X)$ about a random variable X:

$$P_u(x_i) = \frac{1}{n}, x_i \in X = \{x_1, x_2, ..., x_m\} \tag{3}$$

At this point the information entropy is greatest, the uncertainty about X is greatest, and the amount of information implied by X is greatest.

Consider again the deterministic distribution $P_c(X)$ with respect to the random variable X:

$$P_c(x_i) = \delta_{ij} = \begin{cases} 1, & i = j, \\ 0, & i \neq j, \quad j = 1, 2, ..., m \end{cases} \tag{4}$$

At this time, the information entropy is minimized, the uncertainty about X is minimized, and the amount of information contained in X is also minimized to 0. Then, considering the discriminatory information, the amount of information obtained by understanding X from the uniform distribution with equal probability $P_u(X)$ to its true distribution $P(X)$ is $KL(P \| P_u)$, and the amount of information obtained by understanding X from the uniform distribution with equal probability $P_u(X)$ to its fully determined distribution $P_c(X)$ is $KL(P_c \| P_u)$, and then the difference in the amount of information obtained by $P_u \to P_c$ and $P_u \to P$ is:

$$\begin{aligned} & KL(P_c \| P_u) - KL(P \| P_u) \\ &= \sum_{i=1}^{m} P_c(x_i) \log \frac{P_c(x_i)}{1/m} - \sum_{i=1}^{m} P(x_i) \log \frac{P(x_i)}{1/m} \\ &= \sum_{i=1}^{m} P_c(x_i) \log P_c(x_i) + \sum_{i=1}^{m} P_c(x_i) \log m - \sum_{i=1}^{m} P(x_i) \log P(x_i) - \sum_{i=1}^{m} \log m \\ &= -\sum_{i=1}^{m} P(x_i) \log P(x_i) \\ &= H(X) \end{aligned} \tag{5}$$

Thus, for random variable X, the amount of information to know about its equiprobability distribution $P_u(X)$ up to its fully deterministic distribution $P_c(X)$ minus the amount of information to know about its equiprobability distribution $P_u(X)$ up to its true distribution $P(X)$ yields the amount of information about itself, i.e., the information entropy $H(X)$. Formally, the information entropy of random variable X is equal to the difference between the KL scatter of its full distribution with its equal probability uniform distribution and the KL scatter of its true distribution with its equal probability distribution. Since the first item $KL(P_c \| P_u)$ is a definite value, the information entropy is only related to the latter item $KL(P \| P_u)$, which can be interpreted as the larger the KL dispersion

between the true distribution $P(X)$ of a random variable X and its equal probability uniform distribution $P_u(X)$, the smaller its information entropy is and the less information it contains.

2.1.3. Information metrics. For random variable $X=\{x_1,x_2,...,x_m\}$, its a priori uncertainty is its Shannon information entropy [24] $H(X)$, and the random variable with which it has some kind of statistical dependence or dependency is $Y=\{y_1,y_2,...,y_n\}$, then the a posteriori uncertainty about X under the condition of knowing Y is called $H(X|Y)$ also known as conditional entropy. Since the knowledge of X cannot be fully grasped by knowing Y, making $H(X|Y)\le H(X)$, but the a priori uncertainty about X can be reduced by knowing Y, it can also be considered as the amount of information about X obtained by knowing Y. Thus, mutual information is defined as:

$$\begin{aligned} I(X;Y) &= H(X)-H(X|Y) \\ &= -\sum_{i=1}^{m}P(x_i)\log P(x_i)-\left(-\sum_{i=1}^{m}\sum_{j=1}^{n}P(x_i,y_j)\log P(x_i|y_j)\right) \\ &= \sum_{i=1}^{m}\sum_{j=1}^{n}P(x_i,y_j)\log\frac{P(x_i,y_j)}{P(x_i)P(y_j)} \end{aligned} \tag{6}$$

where $P(x_i)$ and $P(y_j)$ denote the prior probabilities of $X=x_i$ and $Y=y_j$, respectively: $P(x_i|y_j)$ and $P(x_i,y_j)$ denote the conditional and joint probabilities, respectively. Mutual information $I(X;Y)$ denotes the uncertainty about X that is reduced by observing or knowing Y. As the degree of correlation (dependence) between X and Y is greater, its mutual information is greater. Mutual information has symmetry i.e:

$$I(X;Y)=I(Y,X)=H(Y)-H(Y|X) \tag{7}$$

Thus, either $I(X;Y)$ or $I(Y;X)$ denotes the amount of information that random variables X and Y provide to each other about each other.

Shannon's information entropy is a measure of uncertainty for a one-dimensional random variable, and its extension to multivariate random variables introduces the concept of joint entropy, e.g., the joint entropy of a binary random variable $(X;Y)$ is defined as:

$$H(XY)=-\sum_{i=1}^{m}\sum_{j=1}^{n}P(x_i,y_j)\log P(x_i,y_j) \tag{8}$$

where $P(x_i,y_j)$ denotes the joint probability. The joint entropy has the following relationship with information entropy, conditional entropy, and mutual information:

$$H(XY)=H(X)+H(Y|X)=H(Y)+H(X|Y) \tag{9}$$

$$I(X;Y)=H(X)+H(Y)-H(XY) \tag{10}$$

Therefore, the relationship between information entropy, joint entropy, conditional entropy, and mutual information is shown in Fig. 1. In the figure, the left and right circles represent the Shannon information entropy $H(X)$ and $H(Y)$ for random variables X and Y respectively, the intersection of the two is the mutual information $I(X;Y)$, and the concatenation of the two is the joint entropy as $H(XY)$.

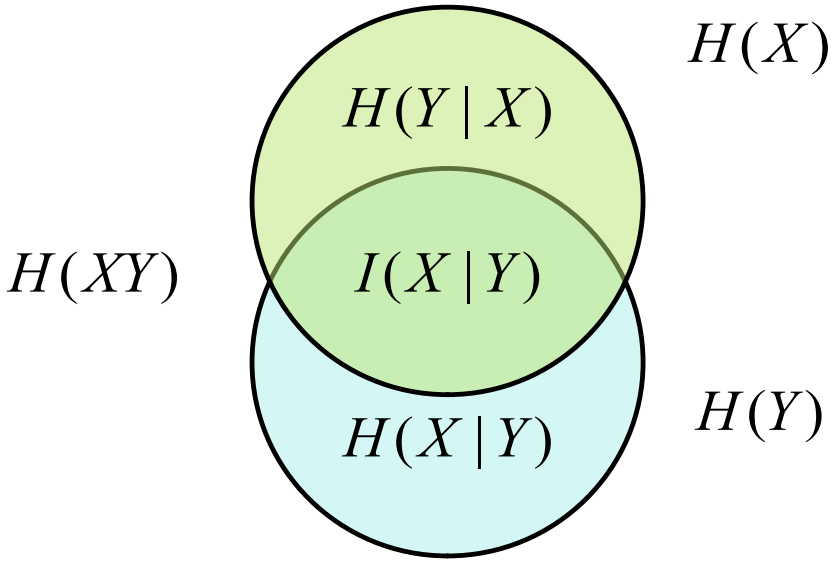

Fig. 1. Information entropy, combined entropy, conditional entropy and mutual trust

In addition to mutual information as an information-theoretic measure of the degree of correlation between two random variables, there are many other forms of information-theoretic measures such as (1) $\frac{I(X:Y)}{H(X)}$ or $\frac{I(X:Y)}{H(Y)}$, (2) $\frac{2I(X:Y)}{H(X)+H(Y)}$, (3) $H(X|Y)+H(Y|X)$, etc., which are derived from the combined forms of information entropy, conditional entropy, mutual information, or joint entropy, and some of which satisfy the three elements of the distance metric, and some of which do not, but all of which are widely used in the specific tasks of data mining.

2.2. Methods for estimating information entropy in foreign language speech

Factors affecting the performance of statistical language models include the order of the model and the basic units used to construct the model. Obviously, the higher the order of the model, the better the performance of the model, but of course, the higher the order of the model, the more difficult to construct. At the same time, the basic unit for constructing the model is also important. The number of modern foreign language vocabulary is too large, and a better lexical corpus has not yet been constructed, so this paper only establishes a binary model of foreign language words. Therefore, the language models realized in this paper are one-, two-, three- and four-dimensional grammar models based on foreign language characters, one-, two- and three-dimensional grammar models based on foreign language words, and binary grammar models based on foreign language words.

When building a language model, due to the irregularity of the training corpus, the number of occurrences of certain linguistic symbol strings $(c_{ik+1}\ldots c_{i+1}c_i)$ may be 0, or $\#(c,_{k+1}\cdots c_{p-1}c_i)=1$ and $\#(c_{i-k+1}\cdots c_{i-1})=1$ in the following equation and lead to $P(c_i\,|\,c_{i+k+1}\cdots c_{i-1}c_i)=1$. This problem is called the data sparsity problem. There are various algorithms to solve this problem, and in this paper, Katz's regressive data smoothing algorithm is used.

Approximate estimation of linguistic information entropy can be realized by the language model established above, and the computational steps for estimating the information entropy of modern foreign languages by using the N-Gram model [25-26] are divided into two steps as follows:

Step 1: Establish a large-scale foreign language corpus, and statistically estimate the n-element co-occurrence probability in the large-scale foreign language monolingual corpus, and estimate the conditional probability in Eq. (11) $P(c_i\,|\,c_{i=k+1}\cdots c_{i-1})$. When the corpus is large enough, according to the method of maximum likelihood estimation and the large number theorem, there are:

$$P(c_i \mid c_{i-k+1} \cdots c_{i-1} c_i) = \frac{\#(c_{i-k+1} \cdots c_{i-1} c_i)}{\#(c_{i-k+1} \cdots c_{i-1})} \tag{11}$$

where and denotes the cumulative number of occurrences of $\#(c_{ik+1} \ldots c_{i,1} c_i)$ and $\#(c_{ik+1} \ldots c_{i,1})$ in the corpus.

Step 2: Approximate $H(P_a)$ using the following formula:

$$H(P_M) \approx -\frac{1}{N}\left(\sum_{i=1}^{n-1} \log_2 P_M(c_i \mid c_1 c_2 \cdots c_{i-1}) + \sum_{i=n}^{N} \log_2 P_M(c_i \mid c_{i-n+1} c_{i-n+2} \cdots c_{i-1})\right) \tag{12}$$

where N is the size of the training corpus. If $P_M(c_i^n)$ is used to denote the probability estimate of model M for c_i^n, equation (12) can be simplified as:

$$H(P_M) \approx -\frac{1}{N}\left(\sum_{i=1}^{n-1} \log_2 P_M(c_i \mid c_1^{i-1}) + \sum_{i=n}^{N} \log_2 P_M(c_i \mid c_{i-n+1}^{i-1})\right) \tag{13}$$

In this paper, taking the large-scale foreign language monolingual corpus covering foreign language news, law, modern official documents, modern literature, foreign language medicine and some religious documents with a total of more than 3 million words as statistical samples, we apply the above formulas to estimate the information entropy of modern foreign languages, and obtain the information entropy and perplexity of modern foreign languages estimated by different language models as follows.

1) The information entropy decreases as the model order rises, which indicates that the more fully the influence of context is considered, the closer the calculation results are to the real value. For example, the higher the model order of foreign language characters, the smaller the information entropy and the greater the redundancy of the model, which reflects the existence of strict constraints between characters.

2) Different models have different degrees of accuracy in language description, the smaller the estimated information entropy or perplexity, the more accurate the model's portrayal of the language.

3) The performance of language models is related to both model units and model orders. The unitary word model is slightly better than the binary word model, but not as good as the ternary word model, and the binary word model is better than the ternary word model.

2.3. Application of structured thinking in foreign language speech interpretation

Structured thinking in foreign language speech interpretation enhances a systematic approach to processing and conveying information, which helps to improve the efficiency and effectiveness of communication. Through this approach, interpreters are able to better organize their thoughts, clearly communicate complex concepts, and ensure that information is transmitted accurately across languages and cultures. The following is an introduction to the application of structured thinking in foreign language speech interpretation:

1) Comprehension and Analysis. Quickly understand the speaker's intention and the gist of the message. Analyze the structure of the content of the speech and identify key and secondary information.

2) Information Transformation. Ensure the accuracy and completeness of the message in the language conversion process. Consider cultural differences and adjust language appropriately to suit the target audience.

3) Logical Maintenance. Maintain the logical structure of the content of the speech to ensure the coherence of the interpreted content. Maintain the hierarchical structure of the information to ensure a clear distinction between the main information and detailed information.

4) Real-time feedback. Adjust the interpretation strategy instantly according to the audience's reaction and the speaker's changes. Be able to respond flexibly when faced with a sudden change of topic or

new information provided by the speaker.
5) Memorization and Recording. Use structured thinking to improve short-term memory capacity and remember key information. Use effective note-taking techniques to record key points in order to interpret accurately.
6) Communication Skills. Pay attention to the speaker's non-verbal messages, such as body language and intonation, which are important for interpreting accurately and conveying emotion. Establish good communication with the audience to ensure that the message is properly understood and received.

3. Interpretation results and analysis of structured thinking in the process of foreign language speech

The experimental corpus for this study consisted of four English speeches, and the speech rate of each speech video was adjusted using the software of Acoustiguide. Based on the developmental stage of the subjects' interpreting ability, a fast version (speech rate of about 140 words/minute, duration of about 11 minutes) and a slow version (speech rate of about 120 words/minute, duration of about 13 minutes) were generated for each speech, totaling eight speech videos. The speech rate was adjusted to keep the intonation unchanged as much as possible, and the two student interpreters who were at the same stage of interpreting ability development as the subjects thought that the adjusted speech had a natural and undistorted voice and intonation after the trial translation. In order to ensure the comparability of the content and style of the speeches, all the speeches were extracted from three live speeches given by the same speaker at different times and on the same occasions. The original speeches have standard pronunciation, moderate speed and difficulty, highly similar themes, and are about 1000 words long.

3.1. Results and analysis of information loss statistics

3.1.1. Results and Analysis of Propositional Information Losses:
1) Overall propositional information loss data. The propositional information loss data include all assertions (T+S+R), the total number of T+S assertions and T assertions, the total number of losses, the propositional information loss rate, the number/rate of untranslated propositions, the number/rate of unformed propositions, and the number/rate of propositional mistranslations for the three types of assertions. The overall data on propositional information loss is shown in Table 1. The results show that the propositional untranslated rate, unformed propositional rate and propositional mistranslated rate in T-assertions are 16.84%, 18.95% and 8.61% respectively. There is only 50% untranslated loss of propositions in R assertions. 8.61% untranslated propositions and 8.57% mistranslated propositions in S assertions.

Table 1. The general data of the propositional information loss

Propositional information loss	Assertion type		
	T	R	S
total	950	30	35
Outstanding number	160	15	9
The proposition is untranslated(%)	16.84	50%	25.71%
Unformed propositions	18		
Unformed propositional rate(%)	18.95%		
Misinterpretation	82		3
Mistranslation rate(%)	8.61%		8.57%

2) Information component loss. The information component loss data include the frequency of each type of information component loss, the number of assertions involved, the frequency assertion ratio and the loss rate. Since there are very few cases of information component loss for R-assertion and S-assertion, only T-assertion loss is counted. If multiple loss types appear in an assertion, one type is counted as being involved in one assertion, two types are counted as being involved in one assertion each, and so on; if the same loss type appears multiple times in an assertion, it is only counted as being involved in one assertion. The graphs are in table space, and the loss types are represented as codes (below). The overall data on information component loss is shown in Table 2. The results show that the frequency assertion ratios and loss rates of T assertions in TFLO, TFLE, TFLS, and TLO loss types range from 3.6%-7.6% and 3.0%-7.1%, respectively.

Table 2. The overall data of the information component loss

Loss type	Loss frequency	Number of assertions	Frequency assertion ratio (%)	Loss rate (%)
TFSPO	6		0.7	
TFSPE	9		1.0	
TSPO	16		1.7	
TSPE	5		0.5	
TFLO	70	62	7.6	6.9
TFLE	68	62	7.3	6.7
TFLS	70	64	7.6	7.1
TLO	32	28	3.6	3.0
TLE	1		0.1	
TLS	11		1.3	

3) Overall information loss. Based on the propositional information and information component loss data, the overall information loss data can be obtained. In view of the large impact of propositional information loss on source language information transfer, the number of propositional information loss and the number of assertions involved in serious loss of information components are combined to obtain the total number of serious loss assertions, which is divided by the total number of assertions to obtain the total serious loss rate. The total number of severe loss assertions and the number of assertions involved in non-severe loss of information components were combined to yield the total number of loss assertions, divided by the total number of assertions to yield the total loss rate. The total number of lossy assertions and the total loss rate reflect the overall loss including propositional information and information components. The overall data on overall information loss is shown in Table 3. The results show that the overall information loss has the highest total loss rate and total severe loss rate among all assertions, which are 50% and 44.5%, respectively; while the non-severe loss rate of information components is the lowest, which accounts for 5.5%; and the propositional

information loss rate and severe loss rate of information components are in the middle of the range, which account for 26% and 19%, respectively.

Table 3. Overall data loss data

Assertion type	T assertion	T+ S assertion	Full assertion
Total claim	950	965	1000
The number of propositional information loss	235	245	260
Propositional information loss rate (%)	24.74	25.39	26
The number of assertions involved in the untranslated, translated, partial translation of the focus information	190	190	190
The information component has a serious loss rate (%)	20	19.68	19
Total loss of information	420	430	445
Gross loss rate (%)	44.21	44.56	44.5
The number of assertions that are not translated, mistaken, and biased	55	55	55
Information component is not a severe loss rate(%)	5.79	5.7	5.5
Total loss assertion total	472	485	500
Total loss rate(%)	49.68	50.26	50

3.1.2. Results and Analysis of Information Loss Data for Each Language Piece:

1) Frequency of information loss by type. The information loss data of the corpus as a whole reflects the overall average loss situation. Because of the different levels of variables in each part of speech, it is necessary to count the information loss data of each part of speech separately. The information loss data of each part of speech includes the frequency of each type of information loss, the frequency assertion ratio, and the loss rate. The frequency of each type of information loss is shown in Table 4. The results show that among the 6 types of propositional information loss, TPO and TPE have a high frequency of information loss, 67 and 30 times respectively, while SPE and SPO are relatively low, 1 and 6 times respectively. Among the 13 types of information component loss, the top three loss types in terms of loss frequency were TFLS, TFLE, and TFLO, whose loss frequencies totaled 55, 42, and 37 times, respectively; while the loss of TLE and SFLO both totaled 1 time.

Table 4. Frequency of information loss in each language

Number		S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
Propositional information loss	TPO	3	5	4	2	6	12	12	4	6	13
	TPE		2	7		2	2	3	2	9	3
	TPI		5	2		2		2	1	1	2
	SPO		2	2						2	
	SPE		1								
	RPO			6		2	2	2			3
Component information loss	TFSPO		3			1		2			
	TFSPE	2	1	3		1	1				
	TSPO		1	2	1	1	1	4	2	2	3
	TSPE		1				2				2
	SSPO	2									
	TFLO	1	13	3	1	4	6	4		2	3
	SFLO									1	
	TFLE	10	11	6	1	4	2	4	1	1	2
	SFLE										2
	TFLS	3	8	3	2	6	3	4	11	5	10
	TLO	3	2	4	3		2		5	5	3
	TLE					1					
	TLS	2	1		2	5					

2) Propositional information loss. The propositional information loss data include the proposition untranslated rate, unformed proposition rate, proposition mistranslated rate, and T-assertion propositional information loss rate for various types of assertions in each language chapter. The propositional information loss data for each part of speech is shown in Table 5. The results show that information loss exists in all discourse parts and the loss types are widely distributed. Taking T-assertion as an example, the proposition untranslated rate of number S7 is the highest (21.8%), and S4 is the lowest (2.7%); the proposition mistranslated rate of S9 is the highest (23.5%), and S5 is the lowest (1.8%); the rate of unformed propositions of S1-S10 is low, ranging from 0-4.9%; and the proposition information loss rate of S9 is as high as 44.5%, followed by S7 (30.3%), and S4 had the lowest rate of 2.8%. From the perspective of the speech parts of speech in each foreign language, information loss is very common, and there are various types of loss.

Table 5. The data loss data of the propositions of each discourse

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
T assertion number	59	58	49	41	67	59	63	49	45	62
The proposition is untranslated (%)	7.1	5.6	6.6	2.7	11	19.7	21.8	11.1	16.4	19.3
Mistranslation rate (%)	10.7	3.8	13.2		1.8	5.6	6.9	6.8	23.5	3.7
Unformed propositional rate (%)		3.8	2.5		1.6		1.9	2.5	4.9	1.8
Propositional information loss rate (%)	17.8	12.9	21.9	2.8	14.4	25.4	30.3	19.8	44.5	24.1
S assertion number	2	8	2							
The proposition is untranslated (%)		51.3	16.9							
Mistranslation rate (%)		100								
R number		3	8		5	2	2			4
The proposition is untranslated (%)		100	58.4		26.2	100	100			66.9

3) Information component loss. Information component loss occurs mainly in T-assertions, so only T-assertion data are counted. The information component loss data includes four categories of frequency assertion ratios: untranslated focus information components (including TFSPO, TFLO), focus information component mistranslation (including TFSPE.TFLE), focus information component bias translation (TFLS), non-focus information component untranslated, mistranslated and biased (including TLO, TLE, TLS), information component severe loss rate and information component non-severe loss rate. Since the frequency of loss of information components and the number of assertions involved in the loss are not exactly equal, the practice of the overall data statistics is still followed here. At the same time, following the previous practice, the ratio of the number of assertions involved in untranslated, mistranslated, and biased translations of focal information components to the total number of assertions is called the severe loss rate of information components, and the ratio of the number of untranslated, mistranslated, and biased translations of unfocused information components to the total number of assertions is called the non-severe loss rate of information components. The data on the loss of information components of T-assertions for each language are shown in Table 6. There are significant differences in the information component attrition data for each part-of-speech T-assertion. There are also significant differences in the frequency assertion ratios of untranslated, mistranslated, and biased translations for the focal information components.S2 has the highest frequency assertion ratio in untranslated and mistranslated with 26.6% and 22.9%, respectively, and the lowest with 3.4% in S9 and 2.7% in S10, respectively, and S8 does not have untranslated. Bias translation was highest at 22.8% for S8 and lowest at 4.7% for S6. The highest rate of severe loss of information components was 41% for S2 and the lowest was only 8.7% for S4.

Table 6. The loss data of the information component of the t assertion

Number	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
The focus information component does not translate frequency assertion ratio (%)	4.6	26.6	5.4	3.7	10.5	13.6	11.2		3.4	4.3
The focus information component error frequency assertion ratio (%)	22.2	22.9	16.3	6.4	8.9	4.7	6.1	5.4	5.8	2.7
The focus information component is the frequency of the assertion ratio (%)	8.1	13.8	9.8	8.2	11.9	4.7	6.1	22.8	10.4	18.6
The non-focused information component is untranslated, incorrect, and the frequency of the translation is the ratio (%)	6.4	6.6	14.1	11.6	12.2	9.8	9.4	11.7	12.7	8.9
The information component is a serious loss rate (%)	27.4	41	24.6	8.7	22.6	18.4	17.5	24.5	12.7	20.2
Information component is not a severe loss rate (%)	4.6	4.7	11.5	8.8	7.3	6.5	6.9	6.8	8.4	6.2

4) Overall information loss. Following what was done for the overall statistics, the propositional information loss rate and the information component severe loss rate for each discourse were combined to give the total severe loss rate. The total severe loss rate and the information component non-severe loss rate were combined to yield the total loss rate. Only T-assertion data were counted. The total loss rate of T-assertion for each part of speech is shown in Table 7. The results show that among the total loss rates of T-assertions for each discourse, S9 has the highest propositional information loss rate and S4 has the lowest, which are 45.3% and 3.8%, respectively. The serious loss rate of information components was highest for S2 (41.2%) and lowest for S4 (8.8%). As for the non-severe loss rate of information components, the percentage of S1-S10 is less than 10%, and the difference between the total severe loss rate and the total loss rate of T-assertions of each discourse is small, and the mean value of both of them in S1-S10 is 42.83% and 49.94%, respectively.

Table 7. The total loss rate of the t assertion

Number	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
Propositional signal loss rate(%)	18.6	13.9	22.8	3.8	15.4	26.3	31.2	20.4	45.3	24.5
The information component is a serious loss rate(%)	27.4	41.2	24.6	8.8	22.4	18.3	17.5	24.6	12.5	20.4
Information component is not a severe loss rate(%)	4.6	4.7	9.3	8.5	7.4	6.6	7.8	7.9	8.1	7.6

Gross loss rate(%)	44.3	53.4	46.8	11.5	37.4	43.2	47.5	44.8	56.1	43.3
Total loss rate(%)	48.4	57.5	57.1	19.3	43.5	49.7	54.5	51.0	63.5	50.9

5) Overall untranslated rate. Following what was done for the overall statistics, the overall untranslated (including unformed propositions), overall mistranslated and overall biased translations were counted for each utterance. Only the T-assertion data were counted. The overall untranslated rate of T-assertions for each language piece is shown in Table 8. The results show that the percentage of untranslated propositions of each language T-assertion is between 20.7% and 24.5% for S6, S7, S9 and S10, and the rest of the numbers are less than 20%. The serious untranslated rate of information components was less than 10% for all numbers except S2 (20.7%); the unserious untranslated rate of information components was smaller for S1-S10, ranging from 0.5%-11.9%.S7 had the highest percentage of total serious untranslated rate and total untranslated rate, with 30.6% and 38.9%, respectively; and the lowest percentage of both for S4, with 4.1% and 12.0%.

Table 8. The total untranslated rate of the t assertion

Number	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
The proposition is untranslated(%)	8.1	10.2	9.8	3.7	13.6	20.7	24.5	14.2	22.1	21.7
The information component is severely untranslated(%)	4.4	20.7	3.2	1.5	8.3	9.6	7.2		1.2	2.1
Gross untranslated rate(%)	9.4	29.8	11.9	4.1	20.8	29.2	30.6	12.0	22.2	22.7
The information component is not a severe untranslated rate(%)	2.4	2.5	11.9	6.8	0.5	4.3	7.2	5.4	6.0	5.2
Total untranslated rate(%)	12.9	33.4	24.9	12.0	22.4	34.6	38.9	18.5	29.2	29.0

3.2. Comparative study of speech information transfer rates

We calculated A number of words, B rate of speech (words per second), C information density (bits per word), and D information rate (bits per second) for each of the eight English speeches in the interpreted material. The histograms and KDE distributions of the four parameters are shown in Figure 2. The results show that in the interpreted material of English speech, 1 time requires an average of 720.55 words, and the rate of narration is an average of 2.25 words per second. According to the calculation results of large-scale text corpus, each language burdens 6.39 bits of information per word on average. Taken together, the languages transmitted an average of 13.45 bits of information per second; in other words, the information rate for interpretation averaged 13.45 bits per second.

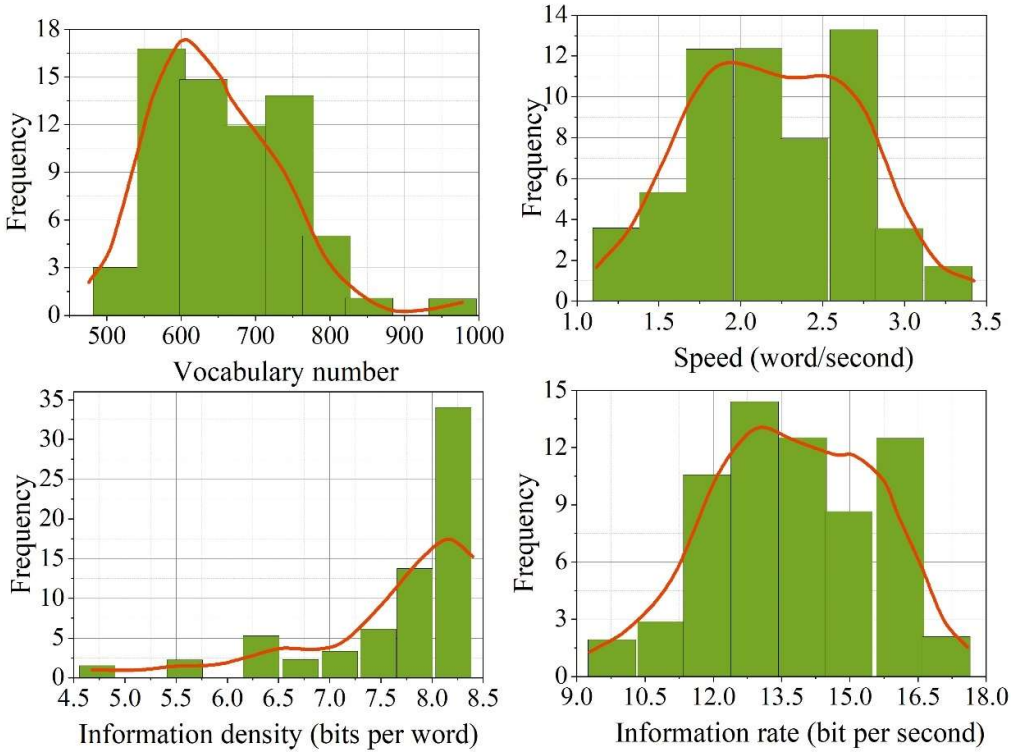

Fig. 2. The histogram of four parameters and the KDE distribution

Next, the differences within each parameter were further examined between the interpreted materials of different foreign language speech materials. The results of pairwise JS scatter for the four parameters sampled are shown in Figure 3. The absolute values of the JS scatter for the three parameters of vocabulary number, speech rate, and information rate are all below 0.078, indicating that the internal variability of each parameter is small. These three parameters are distributed at 0.0655, 0.067, and 0.0705 positions, respectively, and according to the t-test results, there is no significant difference between the distribution of the three two. The information density is distributed on the position of 0.0405, which is significantly smaller than the other three parameters.

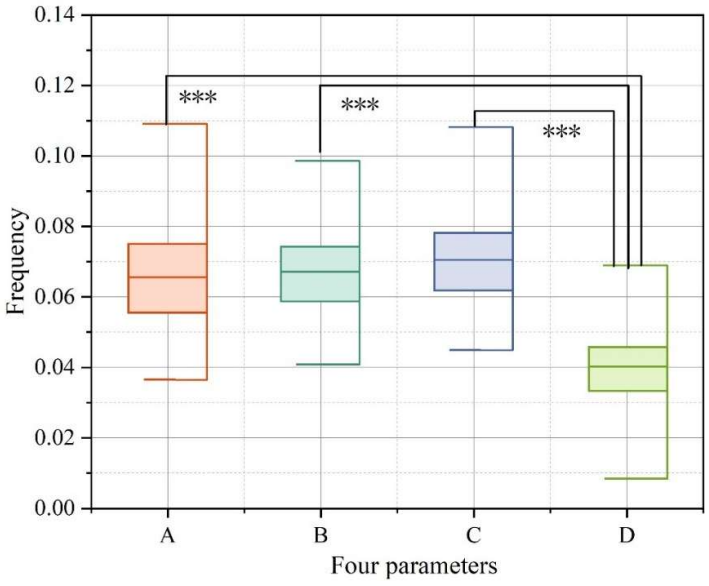

Fig. 3. The four parameters sampled in pairs of JS dispersion results

In order to examine the different characteristics exhibited by the two linguistic units, words and syllables, we calculated the data of the study in syllables and compared them with the results of the present study in words. The results of the pairwise JS dispersion of different parameters of syllable units and word units are shown in Figure 4. The results show that, except for the number of syllables (0.1023), the absolute values of the JS scatter for each parameter range from 0.062 to 0.081, which are all less than 0.09, indicating that all of these parameters in general have very small internal differences. Through comparison, it was found that the variability of the number of words, the rate of speech, and the information density obtained by using words as the unit of calculation was significantly smaller than the corresponding values obtained by using syllables as the unit of calculation for interpreting materials of different foreign language speech materials, but there was no significant difference in the internal variability of the information rate values obtained by both, regardless of whether the unit of calculation was by words or by syllables. The above results indicate that the cross-linguistic variability in information rate is independent of the unit of computation and is comparable and small in the interpreted materials of different foreign language speech materials. However, at the syllable-level unit, the rate of speech and the information density show a large range of variation, making them inversely related, while at the word-level unit, both the rate of speech and the information density are distributed in a relatively stable range, which is significantly smaller than that of the syllable-level unit in all languages.

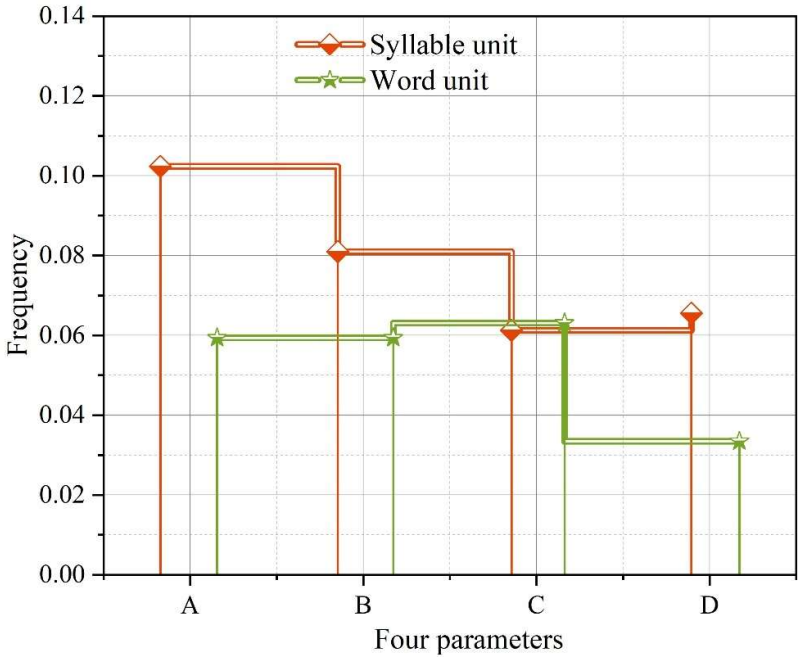

Fig. 4. Paired JS dispersion results for different parameters

3.3. Rater's Evaluation of Subject Interpreter Performance

Two professional simultaneous interpreters watched video recordings of the original speeches and then rated the recordings of the subject's target language output for each of the two speeches. The raters gave separate scores ranging from 0 to 10 on each of the three metrics of completeness, accuracy, and fluency. The average of the two raters' scores was the final score of the subjects. The results of the raters' assessment of foreign language speech interpretation are shown in Table 9. It can be seen that the scores given by the two raters are in good agreement, and the total score error of their ratings ranges from 0.68% to 7.05%. Due to the small sample of data, we did not conduct another statistical validation of the scoring consistency. Overall, the interpretation of the second speech performed better. For specific metrics, subjects generally scored equal or higher in completeness and accuracy for

interpreting the first segment of the speech compared to interpreting the second segment of the speech. From the above data, there seems to be some difference in interpreters' simultaneous interpreting performance when the speaker is directly visible and when the speaker is not visible. Being able to see the speaker's posture seems to have some enhancing effect on the interpreter's performance.

Table 9. The evaluation of the interpretation of the speaker's foreign language speech

Subjects		S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
speech	1	7.2	6.9	7.4	6.9	7.3	7.2	7	6.5	7.1	7.1
	2	7.1	6.9	7.1	7	7.3	7.3	7	6.7	7	7.2
Information integrity	1	7	7.9	7.8	8.2	7.5	8.3	8.1	7.5	7.9	7
	2	8	8.1	8.1	8	8	8.1	8.2	7.4	7.9	7.8
Interpretation accuracy	1	6.9	7	7	6.9	7.6	7.2	7.3	7.2	7.1	7.2
	2	7	7	7.1	7.1	7.1	6.9	7.2	7	7	7
Express fluency	1	6.6	7.5	7.7	7.3	7.4	7.9	7.8	8.2	7	8.5
	2	7.7	8.1	8.2	7.9	7.7	8	8.4	8.1	8	8.1
Total score	1	27.7	29.3	29.9	29.3	29.8	30.6	30.2	29.4	29.1	29.8
	2	29.8	30.1	30.5	30	30.1	30.3	30.8	29.2	29.9	30.1

4. Conclusion

In this paper, the information entropy of modern foreign languages was estimated by using the N-Gram model, and the information loss as well as the rate of information transfer was statistically calculated by using the N-Gram model, and finally the interpretation performance of the subjects was analyzed. The main conclusions are as follows:

1) The statistical results of information loss show that the phenomenon of information loss is prominent, with many types of loss, high frequency, and many cases of serious loss. The results show that there are 8.61%-18.95% propositional information loss, 3.0%-7.6% constituent information loss, and 49.68% overall loss in T-assertions.

2) From the perspective of each speech discourse, the phenomenon of information loss is very common, and there are various types of loss. The data on information loss for each speech segment show that TPO has the highest frequency (67 times) and SPE has the lowest (1 time) among the 6 types of propositional information loss.TFLS has the highest frequency (55 times) among the 13 types of information component loss, and TLE and SFLO have the lowest frequency of loss, which are both 1 time. Among the propositional information loss for T assertions, S7, S4 and S9 corresponded to the highest propositional untranslated rate, propositional mistranslated rate and propositional information loss rate, which were 21.8%, 23.5% and 44.5%, respectively.

References

[1] Rao, P. S. (2019). The importance of speaking skills in English classrooms. Alford Council of International English & Literature Journal (ACIELJ), 2(2), 6-18.

[2] Gwynne, N. M. (2014). Gwynne's Grammar: The Ultimate Introduction to Grammar and the Writing of Good English: Definitions, Explanations and Illustrations of the Parts of Speech, and of the Other Most Important Technical Terms of Grammar. Incorporating Strunk's Guide to Style Explaining how to Write Well and the Main Pitfalls to Avoid. Knopf Publishing Group.

[3] Hashemi, M., & Hokmabadi, M. (2011). Effective English presentation and communication in an international conference. Procedia-Social and Behavioral Sciences, 30, 2104-2111.

[4] Valleva, E. E., & Polukhina, M. O. (2018). Teaching presentation skills in English language classes. Vestnik of Samara State Technical University Psychological and Pedagogical Sciences, 15(4), 55-65.

[5] Guest, M. (2018). Conferencing and presentation English for young academics. Springer Singapore.

[6] Mahdi, D. A. (2022). Improving speaking and presentation skills through interactive multimedia environment for non-native speakers of English. SAGE Open, 12(1), 21582440221079811.

[7] Ohnishi, S., & Ford, J. H. (2015). Student seminar program for improving academic presentation skills for PhD students in science: The effect of language background on outcome. International Journal for Researcher Development, 6(1), 57-76.

[8] Budiyanto, M., Utomo, K. B., & Supriyanto, R. T. (2020). Presentation as One of The Best Models for Teaching Speaking. International Journal of Linguistics, Literature and Translation, 3(8), 179-189.

[9] Zaitseva, N. (2020). Developing English presentation skills as a component of collaboration competence for sustainable development. In E3S Web of Conferences (Vol. 166, p. 10007). EDP Sciences.

[10] Dahan, D. (2010). The time course of interpretation in speech comprehension. Current Directions in Psychological Science, 19(2), 121-126.

[11] Hopkins, J. (2024). Patterns of interpretation: Speech, action and dream. In Sigmund Freud's The Interpretation of Dreams (pp. 123-159). Manchester University Press.

[12] Soyinka, W. (2021). The interpreters. Vintage.

[13] Leanza, Y., Miklavcic, A., Boivin, I., & Rosenberg, E. (2013). Working with interpreters. In Cultural consultation: Encountering the other in mental health care (pp. 89-114). New York, NY: Springer New York.

[14] Dayter, D. (2021). Dealing with interactionally risky speech acts in simultaneous interpreting: The case of self-praise. Journal of Pragmatics, 174, 28-42.

[15] Daminov, N. K., & kizi Yuldoshova, F. M. (2024). CHALLENGES WITH LISTENING AND SPEAKING SKILLS OF FUTURE INTERPRETERS. GOLDEN BRAIN, 2(1), 295-298.

[16] Shandu, T. P. (2014). The pitfalls and pyrotechnics of interpreting live speeches: from church sermons to political speeches. African Journal of Rhetoric, 6(1), 168-194.

[17] Horváth, I. (2017). The speech behaviour of interpreters. Across Languages and Cultures, 18(2), 219-236.

[18] Boggs, E. (2015). Interpreting US public diplomacy speeches (Vol. 75). Frank & Timme GmbH.

[19] Errico, E., & Ballestrazzi, E. (2014). When the speaker is a great performer: A case study on the role of the interpreter in consecutive interpreting. Revista Española de Lingüística Aplicada/Spanish Journal of Applied Linguistics, 27(2), 365-381.

[20] Pöchhacker, F. (2022). Interpreters and interpreting: shifting the balance?. The Translator, 28(2), 148-161.

[21] Xiaoxiao Liu,Yan Zhao,Shigang Wang & Jian Wei. (2024). Mambav3d: A mamba-based virtual 3D module stringing semantic information between layers of medical image slices. Displays102890-102890.

[22] Qiong Liu,Mingjie Cai,Qingguo Li & Chaoqun Huang. (2025). Multi-label feature selection based on adaptive label enhancement and class-imbalance-aware fuzzy information entropy. International Journal of Approximate Reasoning109320-109320.

[23] Aarthi G.,Priya S. Sharon & Banu W. Aisha. (2024). KRF-AD: Innovating anomaly detection with KDE-KL and random forest fusion. Intelligent Decision Technologies(3),2275-2287.

[24] Peter C. Chu & Chenwu Fan. (2017). Exponential leap-forward gradient scheme for determining the isothermal layer depth from profile data. Journal of Oceanography(4),503-526.

[25] Xi Peng,Peng Jia,Ximing Fan & Jiayong Liu. (2025). ENZZ: Effective N-gram coverage assisted fuzzing with nearest neighboring branch estimation. Information and Software Technology107582-107582.

[26] Brahmaleen Kaur Sidhu. (2024). Explore the N-Gram Model of Adaptive Prediction Text. Journal of Educational Research and Policies(9),19-21.