Proceedings of the 9th WSEAS International Conference on Applied Mathematics, Istanbul, Turkey, May 27-29, 2006 (pp638-644)

Statistical Structure of Printed Turkish, English, German, French, Russian and Spanish

ALADDIN SHAMILOV and SENAY YOLACAN
Department of Statistics
Anadolu University
Eskisehir, 26470
TURKEY

Abstract – Interests in the statistical properties of language, the basic tool for communication, has been frequently used for the development of computer sciences such as the construction of efficient binary codes. The language itself may be also regarded as a code for certain conceptual entities. From this point of view, in this study, statistical structures of printed Turkish, English, German, French, Russian and Spanish are examined on the basis of the probability distribution of letters for the same semantic content. Consequently, the optimal language in the sense of coding theory is determined by using Shannon's measure for entropy. During the analysis of the study, we encountered by some known difficulties about the evaluation of Shannon's measure. In order to get over these difficulties, we have established that the regression analysis is a convenient method. So, a regression equation is given for generalization of entropy estimates and related interpretations are given. The main important result of the paper is that the slope of the simple linear regression model gives the approximated value for the entropy of the languages.

Key-words; Entropy of language, Shannon's measure, Optimal language, Semantic content, Coding theory, Regression analysis, ANOVA.

1 Introduction

Information theory is a branch of the mathematical theory of probability and statistics. As such, its abstract formulations are applicable to any probabilistic or statistical system of observations. Consequently, information theory is applied in variety of fields, as are probability and statistics. It plays an important role in modern communication theory, which formulates a communication system as a stochastic or random process [4; 3].

The rationale behind communication systems is the frequency of the letter used in a language. So importance of the language on communication systems can't be denied. Languages are the main tool for the communication between people. So, language itself can be regarded as a code that codes the thoughts [5; 7].

Several papers entend or comment on Shannon's emprical results for English text. Grinetti [8] recalculates Shannon's estimate of the average entropy of words in English text. Burton and Licklider [9] use longer passages of text for Shannon's estimate. Paisley [10] studies entropy variations due to authorship, topic, structure, and time of composition. Treisman [11] comments on contextual constraints in language, and Miller and Coleman [12] provide more data on the entropy of English using Newman and Gerstman's technique.

Shannon's or related estimates are widely used in different applications in [13; 14]. The psychology literature is particullarly rich in entropy estimates.

The important reference works on the subject are the books by Yaglom and Yaglom [15], Cover and Thomas [1] and Weltner [16].

There are several papers such as [5; 17; 18; 19; 20; 21; 22; 23; 24; 25; 26; 27; 28; 29] which present the entropy of different languages. These studies observe the languages separately and without comparisons.

On the opposite of these studies, in this study we decided to make some

comparisons of the Turkish, English, German, French, Russian and Spanish languages considered as alternate codes which carry the same semantic content. The statistical structure of these printed languages are observed on the basis of the probability distribution of letters. Consequently, the optimal language in the sense of coding theory is determined by using Shannon's measure for entropy. During the analysis of the study, we encountered by some known difficulties about the evaluation of Shannon's measure. In order to get over these difficulties, we have established that the regression analysis is a convenient method. So, a regression equation is given for generalization of entropy estimates and related interpretations are given.

2 The Concept of Entropy and Information

The basic concept of entropy in information theory has to do with how much randomness there is in a signal or random event. An alternative way to look at this is to talk about how much information is carried by the signal. The entropy formula expresses the expected information content or uncertainty of probability distribution.

The entropy H(**X**) (Shannon, 1948) of a discrete random variable X is defined by

$$\boldsymbol{H(X)} = -\sum_{x\in\chi} \boldsymbol{p(x)} \log \boldsymbol{p(x)} . \qquad (1)$$

The log is to the base 2 and the entropy is expressed in bits. [1]

In order to reduce uncertainty of system, much important information is required about the system. From this point of view, learning information about the system means reducing the entropy of the system.

Assume that a physical system, say X is given. Let H(**X**) be the entropy of this system. Information of the system means the change of entropy. In the other words, let H(**X**) be the initial entropy of the system and $\boldsymbol{H'(X)}$ be the latter entropy of the system, then the information of the system is defined by,

$$\boldsymbol{I(X)} = \boldsymbol{H(X)} - \boldsymbol{H'(X)} . \qquad (2)$$

If the all information about the system is gained then the entropy of the last system will be $H'(X)=0$. In this case,

$$\mathbf{I(X)=H(X)}. \qquad (3)$$

Hence, it's obvious from (3) that information formula [6] is equal to the entropy formula and defined by

$$\boldsymbol{I_x} = -\sum_{i=1}^{n} \boldsymbol{p_i} \log_2 \boldsymbol{p_i} . \qquad (4)$$

If printed language is indeed an ergodic process, then for sufficiently large *n* a good estimate of H(**X**) can be obtained from knowledge of p(.) on a randomly drawn string ($\mathbf{X_1, X_2, \ldots, X_n}$).

3 Simple Linear Regression

Suppose that one observes n pairs of data, given by $(\boldsymbol{x_i}, \boldsymbol{y_i})$, i=1,2,…,n. The Least Squares problem consists of finding the linear equation given by $\hat{\boldsymbol{y}} = \hat{\beta}_0 + \hat{\beta}_1 \boldsymbol{x}$ which minimizes the following equation

$$\boldsymbol{Q}(\beta_0, \beta_1) = \sum_{i=1}^{n} |\boldsymbol{y_i} - \hat{\boldsymbol{y}}_i|^2 = \sum_{i=1}^{n} [\boldsymbol{y_i} - (\beta_0 - \beta_1 \boldsymbol{x_i})]^2 .$$

β_0, β_1 parameters which minimizes the given function $\boldsymbol{Q}(\beta_0, \beta_1)$ can be found by taking the partial derivatives of $\boldsymbol{Q}(\beta_0, \beta_1)$ with respect to both β_0 and β_1 and setting both of these equations equal to zero. That is,

$$\frac{\partial \boldsymbol{Q}(\beta_0, \beta_1)}{\partial \beta_0} = -2\sum_{i=1}^{n} [\boldsymbol{y_i} - \beta_0 - \beta_1 \boldsymbol{x_i}] = 0$$

and

$$\frac{\partial \boldsymbol{Q}(\beta_0, \beta_1)}{\partial \beta_1} = -2\sum_{i=1}^{n} [\boldsymbol{y_i} - \beta_0 - \beta_1 \boldsymbol{x_i}]\boldsymbol{x_i} = 0.$$

From these equations one obtains the following:

$$n\beta_0 + \sum x_i \beta_1 = \sum y_i$$
$$\sum x_i \beta_0 + \sum x_i{}^2 \beta_1 = \sum x_i y_i$$

The above system of equations is called the system of normal equation for the linear least squares problem. This system has a unique solution given by

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$
$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}},$$

where

$$\bar{y} = \sum_{i=1}^{n} y_i / n \ , \qquad \bar{x} = \sum_{i=1}^{n} x_i / n \ ,$$

$$SS_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - n\bar{x}\bar{y} \ ,$$

$$SS_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum x_i{}^2 - n\bar{x}^2 \ ,$$

$$SS_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum y_i{}^2 - n\bar{y}^2 \ .$$

So, the solution of linear equation is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \ .$$

Achieved results can be also expressed geometrically. Let *X* be independent variable (typically graphed on the horizontal axis), *Y* be dependent variable (typically graphed on the vertical axis). Therefore, we think of *Y* as $f(X)$. The quantities $\hat{\beta}_0$ and $\hat{\beta}_1$ are called *parameters* of the regression equation. The parameter $\hat{\beta}_0$ represents the value of *Y* where the line crosses the vertical axis (that is, $\hat{\beta}_0 = f(0)$), and $\hat{\beta}_1$ is the line's slope.

For linear regression it is assumed that the following hold:

• *Normality:* The dependent data given by $y_1, y_2, \mathrm{K}, y_n \sim N(\mu_y, \sigma_y{}^2)$.
That is the data are normally distributed.

• *Linearity:* $\mu_{\hat{y}} = \beta_0 + \beta_1 x$ .The mean or expected value for y changes as a linear function of x.

• *Homogeneity of variance:* $\sigma_y{}^2$ does not depend upon x.

• *Independency:* The data $y_1, y_2, \mathrm{K}, y_n$ are random. That is, the value for y_i does not depend upon the value for y_{i-1} or y_{i+1} .

If this assumption is violated then the data are said to be correlated or serially correlated.

The regression results are often presented as an analysis of variance table or ANOVA table. The basic idea is to describe how much of the variability found in the dependent variable y can be explained by the presence of the linear equation (($\beta = 0$) versus having a line $y = \bar{y}(\beta \neq 0)$).

4 An Application on Statistical Structure of Languages

The same semantic contents can be described in different ways. For example, in English "creating a low carbon economy", in Turkish "Düşük karbon ekonomisi yaratmak", and in Russian "создание экономики с меньшим содержанием углерода" is the same semantic contents coded in different languages. In other words, the language itself may be regarded as a code for certain conceptual entities. It's obvious that, a comparison can be made in order to obtain the optimal language. The translation from one language to another can be considered as a code transformation. From this point of view in this study, comparisons of the Turkish, English, German, French, Russian and Spanish languages considered as alternate codes which carry the same semantic content are made based on the probability distribution of letters.

In order to obtain the probability distributions of letters of the languages taken into account, a corpus called "Creating a low carbon economy" is chosed as a corpus. The original of the corpus is in English and the translations are in Turkish, German, French, Russian and Spanish [2].

The probabilities of the letters of the different languages are calculated on the basis of computer studies and the related results are given in Table 1 and Table 2.

The same semantic content is coded by 34041 letters based on Russian language which used more letters than the others and by 25065 letters based on English languages which used less letters than the others.

Table 1. The probabilities of Turkish and Russian Letters for the same sample semantic content

Turkish		Russian	
Letter	p_i	Letter	p_i
A	0.0924	А	0.0507
B	0.0165	Б	0.0158
C	0.0093	В	0.0388
Ç	0.0122	Г	0.0142
D	0.0284	Д	0.0268
E	0.0903	Е	0.0780
F	0.0014	Ж	0.0068
G	0.0137	З	0.0125
Ğ	0.0115	И	0.0714
H	0.0046	Й	0.0056
I	0.0507	К	0.0226
İ	0.0844	Л	0.0294
J	0.0009	М	0.0280
K	0.0450	Н	0.0616
L	0.0633	О	0.0855
M	0.0352	П	0.0203
N	0.0575	Р	0.0416
O	0.0180	С	0.0441
Ö	0.0065	Т	0.0534
P	0.0040	У	0.0179
R	0.0641	Ф	0.0025
S	0.0166	Х	0.0103
Ş	0.0175	Ц	0.0035
T	0.0343	Ч	0.0109
U	0.0205	Ш	0.0051
Ü	0.0166	Щ	0.0049
V	0.0106	Ъ.Ь	0.0136
Y	0.0343	Ы	0.0200
Z	0.0145	Э	0.0074
#	0.1250	Ю	0.0049
		Я	0.0169
		#	0.1749

It is assumed that, each conversation in languages is as a population and the chosen corpuses are considered as a random sample selected from the population.

The Shannon's entropy measure is calculated for each corpus based on the languages taken into account. Moreover, total entropy is calculated on the basis of the entropy of the language and the number of letters in the sample. These calculated values are given in the Table 3.

Table 2. The probabilities of English, French, German and Spanish letters for the same sample semantic content

	English	German	French	Spanish
s_i	p_i	p_i	p_i	p_i
A	0.0568	0.0382	0.0526	0.0964
B	0.0136	0.0159	0.0091	0.0114
C	0.0310	0.0199	0.0304	0.0410
D	0.0279	0.0369	0.0349	0.0431
E	0.1182	0.1591	0.1278	0.1207
F	0.0151	0.0175	0.0089	0.0072
G	0.0205	0.0326	0.0112	0.0129
H	0.0265	0.0320	0.0049	0.0030
I	0.0640	0.0758	0.0630	0.0573
J	0.0006	0.0015	0.0017	0.0023
K	0.0049	0.0131	0.0001	0.0000
L	0.0414	0.0277	0.0419	0.0414
M	0.0216	0.0193	0.0231	0.0303
N	0.0655	0.1019	0.0768	0.0648
O	0.0648	0.0228	0.0581	0.0727
P	0.0213	0.0061	0.0274	0.0233
Q	0.0009	0.0004	0.0087	0.0049
R	0.0586	0.0770	0.0666	0.0646
S	0.0513	0.0526	0.0781	0.0681
T	0.0684	0.0496	0.0607	0.0406
U	0.0267	0.0404	0.0507	0.0293
V	0.0110	0.0084	0.0124	0.0086
W	0.0184	0.0187	0.0000	0.0000
X	0.0020	0.0002	0.0039	0.0019
Y	0.0165	0.0002	0.0024	0.0074
Z	0.0002	0.0142	0.0007	0.0028
#	0.1525	0.1180	0.1439	0.1442

Hence, the comparisons of the Turkish, English, German, French, Russian and Spanish languages for communication of same semantic content can be made by interpretations of the Table 3.

According to Table 3, Turkish required 114023 bit, English required 103991 bit, French required 135301 bite, German required 131085 bit, Spanish required 123246 bite and Russian required 147915 bite information for the communication of the same semantic content.

Table 3. The information measures for the same semantic content in Turkish, English, German, French, Russian and Spanish

Translation	Entropy	Number of letters (n)	Total Entropy (TE)
Turkish	4.3299	26334	114023
English	4.1489	25065	103991
French	4.0193	33663	135301
German	4.0796	32132	131085
Spanish	4.0142	30703	123246
Russian	4.3452	34041	147915

Multiple comparisons can be made based on Table 3. Some of them are as follows:

The semantic value of one Turkish letter is %91 of one English letter. In other words, Turkish uses %9 more letters than English to express the same thoughts.

French uses %16 more letters than Turkish.

German uses %13 more letters than Turkish.

Spanish uses %7 more letters than Turkish.

Russian uses %23 more letters than Turkish.

French uses %23 more letters than English.

German uses %21 more letters than English.

Spanish uses %16 more letters than English.

Russian uses %30 more letters than English.

French uses %3 more letters than German.

French uses %9 more letters than Spanish.

Russian uses %8.5 more letters than French.

German uses %6 more letters than Spanish.

Russian uses %11 more letters than Turkish.

Russian uses %17 more letters than Spanish.

Furthermore, in order to generalize the entropy estimates, we select the variate of Total Entropy (TE) as a dependent variable and the variate of number of letters (n) as independent variable. Because, it's obviously seen from Table 3 that the total entropy differs according to the number of letter (in other words total entropy is affected by the probability distribution of letters). Then a simple linear regression model is obtained by MINITAB 13 for Windows as:

$$TE = 6161 + 3.95\, n$$

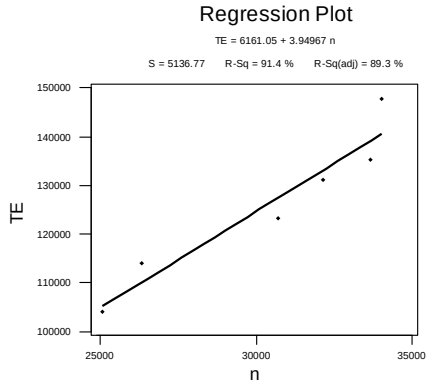
The output of the MINITAB 13 for Windows is as follows:

Predictor	Coef	SE Coef	T	P
Constant	6161	18482	0.33	0.756
n	3.9497	0.6056	6.52	0.003

S = 5136.77
R-Sq = 91.4 %
R-Sq(adj) = 89.3 %

Here P value for t-test statistics is 0.003 so $\hat{\beta}_1$ of the model is significant. The plot of the regression line is shown in Figure 1.

Figure 1. The Regression Plot



Then in order to test the significance of the model, the analysis of variance is made. The results are given in Table 4.

Table 4. ANOVA

Source	DF	SS	MS	F	P
Regression	1	1.122E+09	1.122E+09	42.5402	0.003
Error	4	105545773	26386443		
Total	5	1.228E+09			

According to the Analysis of Variance (AOV) the F statistics is obtained as 42.5402 and the P value for F-test statistics is equal to 0.003. So this regression model is significant for %95 confidence interval. Since R-Sq is equal to 91.4%, it can be said that TE (dependent variable) explains 91.4% of n (independent variable) of the model.

Hence, one can easily estimate the entropy of another language by using mentioned regression model. An interesting result here is that total entropy increases approximately 4 bits per letter whatever the language is.

5 Conclusion

According to the comparisons made in Section 4, it can be said that French and German use almost the same number of letters to express the same thoughts. Turkish uses more letter only than English.

English is the first language that uses the least letters for communication of the same semantic content. In other words, the English letters contains more information than the others.

The important result of the paper is that the slope of the simple linear regression model gives the approximated value for the entropy of the languages.

It must be denoted that, the comparisons can be change according the original of the corpus. For the further studies, also another languages can be taken as original ones and various comparisons can be made.

References:

[1] Cover, T. M. and Thomas, J. A., *Elements of Information Theory*, John Wiley & Sons, Inc, USA, 1991.

[2] DTI, *http://www.dti.gov.uk,* Crown Copyright, 2003.

[3] Hankerson, D., Harris, G. A. and Johnson, P. D., *Introduction to information theory and data compression*, Jr. – 2nd ed. – Boca Raton, Fla.: Chapman & Hall/CRC Pres, 2003.

[4] Kullback, S., *Information Theory and Statistics*, Dover Publications, Inc. ineola, New York, 1997.

[5] Ramakrishna, B. S. ve Subramanian, R., *Relative Efficiency of English and German Language for Communication of Semantic Content*, IRE (IEEE) Transactions on Information Theory; IT-4, **3**, 1958, 127-129.

[6] Shannon, C. E., *A Mathematical Theory of Communication*, Bell System Tech. J. **27**, 1948, 379-423, 623-656.

[7] Yolacan, S., *Statistical Properties of Different Languages Based on Entropy and Information Theory*, Anadolu University Graduate School of Sciences, Master of Science Thesis (at turkish), Eskisehir, 2005.

[8] Grinetti, M., *A note on the entropy of words in printed English*, Inform. Contr., vol.7, 1964, 304-306.

[9] Burton, N. G. and Licklider, J.C. R., *Long-range constraints in the statistical structure of printed English*, American Journal of Psychology, 68, 1955, 650-653.

[10] Paisley, W. J., *The effects of authorship, topic structure, and time of composition on letter redundancy in English texts*, J. Verbal Learn. Behav., 5, 1966, 28-34.

[11] Treisman, A., *Verbal responses and contextual constraints in language*, J. Verbal Learn. Behav., 4, 1965, 118-128.

[12]Miller, G. R. and Coleman, E. B., *A set of thirty-six prose passages calibrated for complexity*, J. Verbal Learn. Behav. 6, 1967, 851-854.

[13] Miller, G. A. and Frick, F. C. , *Statistical behavioristics and sequences of responses*, Psych. Rev., 56, 1949, 311-324.

[14] Siromoney, K. R. and Rajagopalan, K. R., *Style as information in Karnatic music*, J. Music Theory, 8 (2), 1964, 267-272.

[15]Yaglom, A. M. and Yaglom, J. M., *Probability and Information*, 3rd revised ed. Leningrad: Science House, 1973.

[16] Weltner, K., *The measurement of verbal information in psychology and education*, Springer-Verlag, 1973.

[17] Wanas, M. A., Zayed, A.I. Shaker, M. M. and Taha, E. H., *First –second and third-order entropies of Arabic text*, IEEE Trans. Inform. Theory, IT-22 (1), 1976, 123.

[18] Hansson, H., *The entropy of the Swedish language*, Trans. Second Prague Conf. On Inform. Theory, Statistical Decision Functions, and Random Processes, Prague, 1960, 215-217.

[19] Ramakrishna, B. S., Nair, K. K., Chiplunkar, V. N., Atal, B. S., Ramachandran, V. And Subramanian R., *Relative efficiencies of Indian languages*, Nature, 189 (4768), 1961, 614-617.

[20] Küpfmüller K., *The entropy of the German language*, Fernmeldtechnische Zeitschrift, 6, 1954, 265-272.

[21] Petrova N., Piotrovski R., and Giraud R., *The entropy of written French*, Bulletin of Society of Linguistics of Paris, 58 (1), 1964, 130-152.

[22] Manfriono R., *The entropy of the Italian language and its computation*, Alta Frequenza (High Frequency), 29 (1), 1960, 4-29.

[23] Dolezel L., *Prediction of the Entropy and redundancy of Czechoslovakian texts*, Slovo a Sbovesnost, 24 (3), 1963, 165-175.

[24] Nicolau E., Sala C., and Roceric A., *Observations on the entropy of the Rumanian language*, Studisi Cercetai Linvist, 10 (1), 1959, 35-54.

[25] Kazarian R. A., *The evaluation of the entropy of an Armenian text*, News of the Academy of Sciences of Armenia (The Physical and Mathematical Sciences), 14 (41), 1961, 161-173.

[26] Lenskii D. N., *On the evaluation of he Adegei printed texts*, Scientific Notes of the Kabardino-Balkararskii Univ. (The Physical and Mathematical Series), issue 16, Nal'chik, 1962, 165-166.

[27] Ibragimov T. I., *An evaluation of the mutual information of letters in the Tartar language*, Scientific Notes of the Univ. Of Kazan, 124 (2), 1964, 141-145.

[28] Piotrovskaia A. A., Piotrovskii R. G., and Razzhivin K. A., *The entropy of the Russian language*, Questions of Linguistics, 6, 1962, 115-130.

[29] Tarnoczy T., *On factors creating differences in the entropy of a language*, Nyelvtudomanyi Kozlemenyek, 63, 1961, 161-178.