Comparisons of the Turkish, English, German, French, Russian and Spanish Languages For Communication of Same Semantic Content

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

Abstract - The language itself may be regarded as a code for certain conceptual entities. 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. Consequently, the optimal language in the sense of coding theory is determined by using Shannon's measure for entropy and related interpretations are given.

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

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. (Kullback, 1997, Hankerson, D., 2003).

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 (Ramakrishna, B. S. and Subramanian R., 1958; Yolacan S., 2005).

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. Consequently, the optimal language in the sense of coding theory is determined by using Shannon's measure for entropy 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

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

The log is to the base 2 and the entropy is expressed in bits. (Cover and Thomas, 1991)

In order to reduce uncertainty of system, many important information are 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 $H'(X)$ be the latter entropy of the system, then the information of the system is defined by,

$$I(X) = H(X) - H'(X)\,. \tag{2}$$

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

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

Hence, it's obvious from (3) that information formula (Shannon, 1948) is equal to the entropy formula and defined by

$$I_x = -\sum_{i=1}^{n} p_i \log_2 p_i\,. \tag{4}$$

3 Application

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 is in Turkish, German, French, Russian and Spanish (DTI, 2003).

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

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 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

It is assumed that each conversations in languages are 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.

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.

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

Translation	Entropy	Number of letters	Total Entropy
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

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.

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 otherwords, 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.

4 Conclusion

According to the comparisons made in Section 3, 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.

But 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.