Turkish Studies *- International Periodical For The Languages, Literature and History of Turkish or Turkic*
Volume 9/12 Fall 2014, p. 1-7, ANKARA-TURKEY

INTERTEXTUAL COMPARISON OF TURKISH DIALECTS VIA ENTROPY APPROXIMATION*

*Sinan SARAÇLI***

*Cüneyt AKIN****

ABSTRACT

Entropy, a very old term, was developed by Clausius in thermodynamics toward the middle of the 19th century. In thermodynamics or statistical mechanics, entropy is the measure of the level of the existing disorder in a thermo-dynamical system. Today, entropy optimization methods has important applications in the fields such as statistics, mathematics, geography, space sciences, economics, finance, marketing, system analysis, image processing, and model selecting. Furthermore, literature has many works which compare superiority of different languages. There is no such work in the Turkish language. We believe that such future works in the field will fill this gap. Using entropy methods, the present work compares the Göktürk language (the original text), Turkish (translation), Kyrgyz (translation) and Kazakh (translation) in terms of their language characteristics. The MATLAB software was used. An example application was conducted in order to enable the work to be more understandable for the experts working in the field of the Turkish language. Entropy between the texts in the Orkhon inscriptions and their Kyrgyz and Kazakh translations was done. The work used the mutual translations of the same text in the Göktürk language, Turkish, Kazakh and Kyrgyz. Entropy of each text was calculated by the MATLAB software and shown in the tables. The results were obtained by analyzing the text which was used in the application by the program which calculates entropy of any text which was previously entered in the system. The appendix 1 includes the MATLAB codes to calculate eny given text. Any researchers can calculate the entropy of their text through the software by entering their own text.

Key Words: Entropy, Inter-text Comparisons, Turkish Language

---

* *This is an extended and unabridged version of the paper presented at TCSSE International Conference of Social Science and Education 4-6 August 2014, Cornell University, New York USA.*

Bu makale Crosscheck sistemi tarafından taranmış ve bu sistem sonuçlarına göre orijinal bir makale olduğu tespit edilmiştir.

** Yrd. Doç. Dr. Afyon Kocatepe University, Faculty of Arts and Sciences, Department of Statistics, Afyonkarahisar-Turkey El-mek: ssaracli@aku.edu.tr

*** Yrd. Doç. Dr. Afyon Kocatepe University, Faculty of Arts and Sciences, Department of Modern Turkish Dialects, Afyonkarahisar-Turkey El-mek: cuneytakin@aku.edu.tr

ENTROPİ YAKLAŞIMIYLA TÜRK LEHÇELERİ ÜZERİNDE METİNLERARASI BİR KARŞILAŞTIRMA

ÖZET

Çok eski bir kavram olan Entropi ilk olarak termodinamikte 19. Yüzyılın ortalarına doğru Clausius tarafından geliştirilmiştir. Termodinamikteki ya da istatistiksel mekanikte entropi bir termodinamiksel sistemde var olan düzensizlik düzeyinin bir ölçümüdür. Günümüzde entropi optimizasyon metotları, istatistik, matematik, coğrafya, uzay bilimleri, ekonomi, finans, pazarlama, sistem analizi, görüntü işleme, model belirleme gibi alanlarda önemli uygulamalara sahiptir. Ayrıca literatürde, farklı dillerin birbirlerine göre üstünlüklerinin karşılaştırıldığı birçok çalışmaya da rastlanabilir. Türk dili alanında rastlamadığımız bu tür çalışmalar, bu alanda bir boşluğu dolduracaktır kanaatindeyiz. Dile sayısal bir yaklaşım amacıyla, Entropi alanının metotlarıyla gerçekleştirdiğimiz bu çalışmada, aynı düşünceyi ifade eden Göktürkçe (asıl metin), Türkiye Türkçesi (çeviri), Kırgız Türkçesi (çeviri), Kazak Türkçesi (çeviri) metinler dil özellikleri bakımından istatistik biliminde önemli bir yere sahip olan Entropi yaklaşımı ile metinlerarası bir karşılaştırılmaya tâbî tutulmuştur. Çalışmada karşılaştırmayı gerçekleştirmek için MATLAB yazılımından yararlanılmış ve çalışmanın Türk dili alan uzmanları açısından daha anlaşılır olmasını sağlamak amacıyla bir örnek uygulama da yapılmıştır. Orhun abidelerindeki metinler ile, Kırgız Türkçesi ve Kazak Türkçesi aktarımları arasında bir entropi uygulaması gerçekleştirilmiştir. Çalışmanın uygulama kısmında **Göktürkçe,** Türkiye Türkçesi, Kazak Türkçesi ve Kırgız Türkçesi lehçelerinde aynı metinlerin birbirine çevirileri kullanılmış ve MATLAB paket programı yardımı ile her bir metnin Entropisi hesaplanarak tablolarda gösterilmiştir. Çalışmanın uygulama kısmında kullanılan metin, daha önceden sisteme girilen herhangi bir metnin entropisini hesaplayan bir program aracılığı ile analiz edilerek ilgili sonuçlara ulaşılmıştır. Ek 1. verilen herhangi bir metnin entrpisini hesaplamak için MATLAB kodlarını içermektedir. Dileyen bir araştırmacı bu yazılımı elde ettikten sonra kendi metnini girerek bu metnin entropisini hesaplayabilir.

Anahtar Kelimeler: Entropi, Metinlerarası Karşılaştırma, Türk dili

1. Introduction

In terms of language history, the Göktürk (Orkhon) inscriptions, the oldest known text written by Turks, belongs to the first period of the Old Turkish. Thus, the language in these documents represents the oldest known form of the Turkish language. Comparing that language with the examples of its later phases (Kazakh, Kyrgyz and Turkish) through the present work, we believe, is important in terms of measuring the power of expression and efficiency in transmitting information. For the Old Turkish, the translation of Muharrem Ergin, and for the other dialects, the inscriptions' translations to those languages were used in the present work.[1]

Although there were entropy works comparing Turkish with other languages[2], no work which used this method between Turkish and its dialects was found. The first work known which

evaluates the statistical characteristics of the Turkish language was developed in the context of "Armed Forces Reading and Writing Project" by Turkish Ministry of Education, Turkish Ministry of Defense, American Committee of Assistance, American Committee of Financial Assistance and Georgetown University. In the project, Pierce (1963) counted words using a mechanical sequencer on verbal and written texts. Atli (1972) worked on a text containing about 34.000 words. In his master's thesis, Töreci (1974) worked on a text containing 22.216 words. His work dealt with the frequency of characters and letters separately in roots, trunks and words, the frequency of letters in specific places of words, the frequency of letters in two-fold and three-fold forms, the possibility of two-conditional and three-conditional forms, word patterns in vowels and consonants, the distribution of words in terms of the numbers of vowels and consonants, the accordance of vowel harmony when two, three or four vowels come sequentially, the frequency of vowels in two-fold and three-fold forms, the frequency of syllable types, the frequency of syllables, how often syllable types comes after each other, the distribution of words in terms of the numbers of syllables and letters, the distribution of first, middle and last syllables among themselves, and the numbers of roots, trunks and words separately (Gönenç, 1976; reported by Yolaçan, 2005).

Shannon stated that a language's possibility of letters, both conditional and together, may be used as a model of that language. Based on that view, Töreci (1978), in another related work, obtained some of the text characteristics of Turkish until the second level by using the possibility of conditional letters of Turkish. Then the researcher reproduced text examples of Turkish until the ninth level through a computer program which he wrote for that purpose (Yolaçan 2005).

In a recent related work, Yolaçan (2005) compared different coding techniques in the same text written in Turkish, English, French, German, Spanish and Russian, and then calculated the entropy value of the example text for each language. The work on the text expressing the same thought in different languages revealed that Turkish used more letters from English only. It used fewer letters from French, German, Spanish and Russian. It was emphasized that Turkish came second after English in terms of importance.

1. Material and Method

Historically, many notations of entropy have been proposed. The etymology of the word entropy dates back to Clausius (Clausius 1865), who dubbed this term from the greek tropos, meaning transformation, and a prefix en- to recall the indissociable (in his work) relation to the notion of energy (Jaynes 1980). A statistical concept of entropy was introduced by Shannon in the theory of communication and transmission of information (Lesne, 2011). In information theory, entropy is a measure of the uncertainty in a random variable. In recent studies it can be seen that applications of entropy takes place in almost every brunch of science. While there are different kinds of methods in entropy, the most common maximum entropy (MaxEnt) method maximizes the Shannon's entropy (Çiçek, 2013).

For a random variable X, Shannon's maximum entropy H(x) can be calculated as in equation 1.

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

Letters in text that are given in the application part of this study, correspondance to the x random variable given in equation 1. The maximum entropy value, *H(x)* can be calculated after calculating the probabilities of each letters used in the texts. The sign minus in front of the equation is just because to obtain a positive result. (It's known that the logaritms of the probabilities which are between 0 and 1 will give a negative result).

An Illustrative Example

Two texts which have the same meaning and written in English and Turkish are given below.

"***an application on entropy of languages***" and

"***dillerin entropisi üzerine bir uygulama***"

Examining these two texts may help to us to understand the calculation of entropy more celearly. Descriptive statistics for two of these sentences are given in Table 1.

Table 1. Descriptive statistics for two sentences writen in English and Turkish.

Text in English (33 letters and 5 spaces)			**Text in Turkish** (35 letters and 4 spaces)		
Letter	**Frequancy**	**Probability**	**Letter**	**Frequancy**	**Probability**
Space	5	0.132	Space	4	0.103
a	5	0.132	a	2	0.051
c	1	0.026	b	2	0.026
e	2	0.053	d	1	0.026
f	1	0.026	e	4	0.103
g	2	0.053	g	1	0.026
i	2	0.053	i	6	0.154
l	2	0.053	l	3	0.077
n	5	0.132	m	1	0.026
o	4	0.105	n	3	0.077
p	3	0.079	o	1	0.026
r	1	0.026	p	1	0.026
s	1	0.026	r	4	0.103
t	2	0.053	s	1	0.026
u	1	0.026	t	1	0.026
y	1	0.026	u	2	0.051
			ü	1	0.026
			y	1	0.026
			z	1	0.026

Probabilities of each letter given in Table 1. are calculated by dividing the frequency of the letter by the number of the letters in the sentences. As an example when we calculate the probablity of letter "a" in Turkish text, the frequancy of this letter (it is equal to 2) is divided by the number of the letters used in the text (it's equal to 39) and the result is obtained as: 2/39=0,051.

Calculation of the Shanon's maximum entropy, H(X) for the text given in English, by the help of Table 1 we can obtain the result as:

$$H(X) = -[(0.132\log_2 0.132) + (0.132\log_2 0.132) + (0.026\log_2 0.026) + (0.053\log_2 0.053) + (0.026\log_2 0.026) + (0.053\log_2 0.053) + (0.053\log_2 0.053) + (0.053\log_2 0.053) + (0.132\log_2 0.132) + (0.105\log_2 0.105) + (0.079\log_2 0.079) + (0.026\log_2 0.026) + (0.026\log_2 0.026) + (0.053\log_2 0.053) + (0.026\log_2 0.026) + (0.026\log_2 0.026)]$$

$$H(X) = -[(-0.385) + (-0.385) + (-0.138) + (-0.224) + (-0.138) + (-0.224) + (-0.224) + (-0.224) + (-0.385) + (-0.342) + (-0.289) + (-0.138) + (-0.138) + (-0.224) + (-0.138) + (-0.138)]$$

H(X) = -[-3.73256] and **H(X) = 3.73256**

Same as, for the text given in Turkish, Shanon's maximum entropy can be calculated as:

$H(X) = -[(0.103\log_2 0.103) + (0.051\log_2 0.051) + (0.026\log_2 0.026) + (0.026\log_2 0.026) + (0.103\log_2 0.103) + (0.026\log_2 0.026) + (0.154\log_2 0.154) + (0.077\log_2 0.077) + (0.026\log_2 0.026) + (0.077\log_2 0.077) + (0.026\log_2 0.026) + (0.026\log_2 0.026) + (0.103\log_2 0.103) + (0.026\log_2 0.026) + (0.026\log_2 0.026) + (0.051\log_2 0.051) + (0.026\log_2 0.026) + (0.026\log_2 0.026) + (0.026\log_2 0.026)]$

H(X) = -[(-0.337) + (-0.22) + (-0.136) + (-0.136) + (-0.337) + (-0.136) + (-0.415) + (-0.285) + (-0.136) + (-0.285) + (-0.136) + (-0.136) + (-0.337) + (-0.136) + (-0.136) + (-0.22) + (-0.136) + (-0.136) + (-0.136)]

H(X) = -[-3.92593] and **H(X) = 3.92593**

Information of a text has a negative correlation with the entropy (amount of the uncertinity) of a text and these results indicate that the information obtained from the text written in English is more than the information of the text written in Turkish which means English is superior than Turkish according to this comparison.

2. Application

In the application part of this study, the texts corresponding the same meaning but written in Turkish dialect, **Gokturk** dialect, Kırghız dialect and Kazhak dialect are examined via MATLAB software. Entropy values of each dialect calculated as presented in the illustrative example and the results are given in Table 2. MATLAB codes that calculate the entropy of any given text are also given in Appendix 1.

Table 2. Descriptive Statistics and Entropy Values for the Texts Written in Different Dialects

	Texts written in	Number of Letters	Entropy
South Side	Turkish Dialect	2638	4.5743
	Gokturk Dialect	2346	4.5563
	Kırghız Dialect	2397	4.5477
	Kazhak Dialect	2554	**4.4575**
East Side	Turkish Dialect	2295	4.5064
	Gokturk Dialect	2448	4.5001
	Kırghız Dialect	2252	4.4887
	Kazhak Dialect	2456	**4.3795**

Because entropy is the measure of the uncertinity, the maximum information can be obtained from the the dialect that has the minimum entropy value and as a result the dialect which has the minimum entropy can be considered as the superior one.

Table 2 shows that entropy of Kazhak dialect for the south side, is the lowest with the score of 4.4575 which is an indicator that this dialect is the best of all. The entropy amounts of Kırghız Dialect, **Gokturk** Dialect and Turkish Dialects are 4.5477, 4.5563 and 4.5743 resepectively.

For the east sides of the dialects again Kazhak dialect has the lowest entropy with the score of 4.3795 and entropy amounts of Kırghız, **Gokturk** and Turkish Dialects are 4.4887, 4.5001 and 4.5064 resepectively.

Results given in Table 2 also shows that information obtained from the east sides of the dialects is more than the information obtained from the south sides. On the other hand there is no relation between the number of the letters used in the texts and the calculated entropy values. This is an important result because one can think that entropy value increases or decreases by an increasing number of the letters used in the texts.

3. Results and Discussion

There are many different studies examining the languages via entropy approximation. In this study the texts which have the same meaning and written in Turkish dialect, **Gokturk** dialect, Kırghız dialect and Kazhak dialect are examined statistically according to their own linguistic structures.

Results of the study indicate that according to South side of the texts, Kazhak dialect has the minimum entropy among others which means that this dialect is the best of all as a linguistic structure. Kırghız Dialect, **Gokturk** Dialect and Turkish Dialect have the minimum entropies respectively. The same results are also acceptable for the East side of the texts.

Like other studies conducted about structures of the languages via entropy approximation, we hope that this study will help the researchers who are interested in linguistic researches.

REFERENCES

ATLI E., (1972), Yazılı Türkçede Bazı Enformatik Bulgular, Uygulamalı Bilimlerde Sayısal Elektronik Hesap Makinalarının Kullanılması, 409-425, TBTK, Ankara.

BAZILHAN, N., Kazakstan Tarihı Turalı Türki Derektemeleri (2005), R. B. Süleymanov Doğu Bilimleri Enstitüsü Yayınları, Almatı, Kazakistan.

ÇIÇEK, H. (2013), Maksimum Entropi Yöntemi İle Türkiye'deki Coğrafi Bölgelerin Yıllık Hava Sıcaklık Değerlerinin İncelenmesi, Yüksek Lisans Tezi, Afyon Kocatepe Üniversitesi, Fen Bilimleri Enstitüsü, Afyonkarahisar.

ERGIN, M., (1999), Orhun Abideleri, Boğaziçi Yay., İstanbul.

GÖNENÇ G., (1976) Doğal Diller alanında Bilgisayar kullanımı, Bilişim'76 Bildiriler, Türkiye Bilişim Derneği Yayınları, 3(28.1-28.10).

KANTAR M. Y. (2006). Entropi Optimizasyon Metotlarıyla Rassal Değişkenlerin Dağılımlarının İncelenmesi, Doktora Tezi, Anadolu Üniversitesi, Fen Bilimleri Enstitüsü, Eskişehir.

PIERCE, J.(1963), Türkçe Kelime Sayımı – A Frequency Count of Turkish Words, Milli Eğitim Bakanlığı Yayım Müdürlüğü, Ankara.

SIDIKOV, S., KADIRALI, K., Bayırkı Türk Cazuusu VII-X asırlar (Eski Türk Yazısı), Bişkek, 2001.

TÖRECI E.,(1974), Statistical Investigation on the Turkish Language Using Digital Computers, Yüksek Lisans Tezi, ODTÜ Elektrik Mühendisliği Bölümü, Ankara.

DSOUZA, C., (2012). Compute the Entropy of an entered text string, (Source: http://www.mathworks.com/matlabcentral/fileexchange/38295-compute-the-entropy-of-an-entered-text-string)

TÖRECI, E.,(1978), Özdevimsel Olarak Türetilen Türkçe Metin Örnekleri, Bilişim'78 Bildiriler, 166-174, Ankara.

WU X. (2003). Calculation of Nex-Entropy Densities with Application to Income Distribution, Journal of Econometrics **115:** 347-354.

YOLAÇAN Ş. (2005), Farklı Dillerin Entropi ve İnformasyon Teorisi Açısından İstatistiksel Özellikleri, Yüksek Lisans Tezi, Anadolu Üniversitesi Fen Bilimleri Enstitüsü, Eskişehir.

Appendix 1. MATLAB Codes That Calculates the Entropy of a Given Text**.**

(Source: Dsouza, 2012).

```
clear all

s= "ENTER THE TEXT HERE"

H=ComputeEntropy(s);


function H = ComputeEntropy(s)
if (ischar(s)==1)     % Checks whether s is a character array
l=length(s);
uniqueChars = unique(s);    % String s has all unique characters sorted
lenChar=length(uniqueChars);
f=zeros(1,lenChar);
for i=1:lenChar
    f(i)=length(findstr(s,uniqueChars(i)));  % Count the occurence of
unique characters
end
p=zeros(1,lenChar);
for i=1:lenChar
    p(i)=f(i)/l;        %  Probabilities for each unique character
end
H=0;
for  i=1:lenChar
    H = H + (-p(i)*log2(p(i)));  % Calculating the Entropy
end
else
    display('Invalid String');
end
```