Yılmaz, Ç. (2021). An Insight into the Coding Process in the Natural Languages in Terms of Information Theory, International Journal of EurasiaSocialSciences (IJOESS), 12(43), 88-113.

DOI: *http://dx.doi.org/10.35826/ijoess.2836*

ISSN: 2146-1961

ArticleType: *Research Article*

AN INSIGHT INTO THE CODING PROCESS IN THE NATURAL LANGUAGES IN TERMS OF THE INFORMATION THEORY (IN THE EXAMPLE OF TURKISH)

Çağlayan YILMAZ
Asst. Prof, Artvin Coruh University, Artvin, Turkey, cagom49@yahoo.com
ORCID: 0000-0002-1995-212X

Gönderim tarihi: 02.12.2020 Kabul tarihi: 28.02.2021 Yayım tarihi: 15.03.2021

ABSTRACT

This study focuses on the transformation that morphemes undergo in the coding process. This transformation is related to the use. Morphemes marked according to their articulation evolve according to the frequency in the messages. In the study, a method, which will be used in the cases that the Turkish morphemes are coded according to the information of value of Turkish morphemes. This method is based on the information theory which is developed by C. E. Shannon and which describes "meaning" as a measurable concept. This method relates the information value of symbols in the messages to their frequency in the messages. Therefore, it was focused on the relation between the frequency and their code numbers by giving several examples. The information of value of the several Turkish words is measured depending on their frequency and the code number which is necessary for the coding of each morpheme, depending on these values. However, it is necessary that a corpus should be formed and the separation of each text into morphemes and that the frequency of each morpheme within the corpus should be determined. At that stage, the data formed by the analysis of the text pieces and composed of 100 text pieces published specifically in the last decade and which have been prepared by the author of this study for a previous study. In addition, the entropy of the Turkish morphemes has been calculated by using their frequency in the corpus and, the relationship between the number of letters in the alphabet and the number of signs required to mark the words is presented. Finally, the preferred method in the study is presented for discussion. The preference between the required measurement method and the possible measurement method is at a dimension which will require the questioning of the information theory.

Keywords: Informational theory, entropy, coding, alphabet.

INTRODUCTION

When "system" is considered as a whole doing a work, it is no longer possible to define it as a combination of random parts. In this case, each part of the system is the "element" that takes charge in the process of performing this work. Elements work in a cooperation and they are defined according to these roles. The primary function of "language" as a system is to generate "messages" for "communication". The elements of the language to perform this function are "signs" (İmer et al., 2011: 142). Signs, which are the smallest meaningful parts in the messages, should also be defined according to their function in the language system. Each element gets its value according to its position within the system (Barthes, 1979: 52).

When element is defined as the functional part in the system, then its materialistic characteristics are not privileged. What is important is the functional characteristics of the mentioned parts. System is not the sum of materialistic characteristics but it is defined as the cooperation of the parts under discussion (Piaget, 1982: 14). As a result of this approach, the parts performingthe same function become the parts of the same element even if their materialistic characteristics.

In fact, the language system consists of signs that are interchangeable (or changeable) parts, allowing messages to be decoded and reconstructed. Otherwise, no message would be understood, decoded, and consequently reproduced (Lévi-Strauss 2013: 45). Messages that can be decoded and reproduced play a decisive role in the evolution of the language system. Saussure's opposition to "language – parole" is actually a distinction that points to this evolution (Saussure, 1998: 50). Because this opposition ("language – parole" opposition) relates the meaning of the sign to its use in messages.

Language signs initially form messages through the auditory channel. The visualization of these messages can be examined in two interrelated stages in the cultural development of humanity. Even if it has been proposed for a different purpose, the distinction that Martinet calls "double articulation" (Martinet, 1998: 25), forms the basis of both stages in the visualization process of auditory signs. Although the process that evolved from visualizing the signs as a whole (first articulation) to being broken down and coded according to their articulation (second articulation) is observed in indistinct transitions in every culture, coding the signs in text is an interesting story from logograms to graphemes.

The visualization of "morphemes" and "number" as a sign has gotten a new dimension with the use of "letters" and "digits". For the coding process of morphemes, Faulmann's "Yazı Kitabı – Tüm Yerkürenin, Tüm Zamanların Yazı Göstergeleri ve Alfabeleri" (The Book of Scripture Contains the Characters and Alphabets of All Times and of All the Peoples of the World) and for the oding of numbers Ifrah's "Rakamların Evrensel Tarihi" (The Universal History of Numbers) may be used. It would be a necessity to use a different code for each "morpheme" and for each "number" without them. Coding morphemes and numbers by dividing them into elements is a choice of method resulting from the inevitable increase in the number of signs. For this, first of all, signs should be examined and classified as qualities.

It seems that an important distinction seems to be made between morphemes and numbers in this respect just at the beginning. While numbers are separated into their "orders" by making a quantity calculation, morphemes are analyzed according to their characteristics of articulation within this "articulation". In a clearer expression, "value" in numbers and "linearity" in morphemes are taken into consideration while elementary parts are determined. "Linearity" in linguistics is a concept proposed for emphasizing the concatenation of linguistic signs to the time in the articulation process (Saussure, 1998: 115).

The separation of morphemes into elementary codes as independent of the content of morphemes (Martinet, 1998: 206) is indeed a necessary fiction for the beginning level of repeated permutation. Each repeated permutation will get away from the fiction which has created it at the beginning in the "communication" process because the developments in human’s social life will make communication more complex every single day and, hence, the sign number will increase in an unpreventable way. In this way, "messages" created for making communication possible will increasingly require more signs and this will bring the increased in time, energy and cost increase in wake. It is necessary to search for the evolution of writting here.

First of all, the purpose of the study should be clearly stated. It must be struck that the goal in all scripts from the first known attempts to modern times, is to visualize the morphemes used in an auditory state in language, the difference between caption and graphic text is only the difference in method. At first glance, dividing the morphemes into pieces and marking each piece seems more useful than specifying a code for each morpheme (considering the number of morphemes in languages) to reduce the number of codes in the alphabet.

This situation can be made concrete with an example. Let there be a language with 500 morpheme in it. If a "code" is used for each morpheme, an alphabet of 500 codes must be learned. However, if each morpheme is coded by dividing it into 2 parts based on the second articulation of the language, there should be 23 codes in the alphabet (Since $23 \times 23 = 529$, 29 of the code is meaningless). Each morpheme can also be coded by dividing it into 3 parts. In this case, 8 codes in the alphabet are sufficient (Since $8 \times 8 \times 8 = 512$, 12 of the code is meaningless).

As can be seen, the second articulation primarily reduces the number of codes in the alphabet, thus easing the learning by relaxing the memory. Of course, in natural languages, not all morphemes are written with the same number of codes as above. This is the primary question that needs to be answered. That is, a method is required to determine how many codes a morpheme should be written in, independent of its auditory dimension. In this case, the codes used when writing a morpheme cannot correspond to parts of the sound image of that morpheme.

On the other hand, while the number of codes in the alphabet decreases, the number of codes in the texts increases and this causes more fatigue. Continuing with the previous example, when a text consisting of 5000 morphemes is written with an alphabet in which each morpheme is encoded with 1 code 5000 codes are in the text, 10000 codes in the text when each morpheme is encoded with 2 codes, and 15000 codes in the text when each morpheme is written with 3 codes.

As it can be understood, human beings are looking for a coding method that they can learn and write more easily. For this reason, all writing strings (consciously or unconsciously) designed to be learned more easily evolve to be written more easily. In order to analyze the writing systems that have evolved with the tendency to write a text with the least number of codes without increasing the number of codes in the alphabet (Gemalmaz 1982: 28), it is necessary to pay attention to the relationships they establish with each other in the system and paradigm. Therefore, initially, the auditory dimension of morphemes should be excluded from the study. Although the change of a sound image depends on the function of the moprfemin that the sound image corresponds to in the system, associating the codes used to write morphemes with the articulation of the sound images of these morphemes may be accomplished at a later stage.

METHOD

This method, which associates "meaning" (informational value) with frequency, explains indeed how the relation between "signifier" and "signified" ("symbol" and "the block code which is the equivalent of the symbol") becomes arbitrary because the causality situation which exists relatively at the beginning disappears as dependent on frequency. Then, it is necessary to see a sign with the value that appears as a result of the use as a "culture" product. The reason for the arbitrariness of sign lies here. As it is in every culture institution, this is very important in terms of Saussure's "language-word" coincides with the opposition "an abstract permutation-a concrete product". Word is a product of language which is an abstract permutation. Language becomes concrete through word and evolves by means of word (Saussure, 1998: 50). In a clearer expression, it is expected that signification (more precisely, the fact that there is a signifier of a signified) appears as the frequency of sense data, which can be articulated in every way and which are used for communication, increases in language that is a culture institution.

The signification process dependent on the use in Turkish which is a cultural permutation can surely be observed in a "corpus" formed of messages seen in different communication environments. However, a corpus was not prepared in this study and the corpus in a previous study (Yılmaz, 2018) was used. A corpus formed of the parts quoted from the texts, which were published especially in the last ten years and which were mostly close to the colloquial language, (narratives such as story, novel, fairy tale, memoir, joke; news and critical writings about politics, culture, religious and political discourse in the newspapers and journals; theatre and plays; idioms, proverbs, and poems). This was a corpus composed of 100 text pieces (2623 sentences in total), each of which was 26 sentences in average (Yılmaz, 2018: VI).

In this part of the study, the morphemes of Turkish will take place of symbols used in messages. Martinet uses the term "morpheme" in the way it means "the smallest unit particular to the permutation called language" instead of "signification" (Martinet, 1998: 23). However, this definition for the term "morpheme" should be expanded by considering the structure of the language in question. As it is known, Turkish is an agglutinative language. The "words" are formed by adding "affix" morphemes (in a better expression, "suffix") to the "root" morphemes. The roots keep their articulation characteristics relatively more compared to affixes in this combination. Affixes adapt to the articulation of the previous morpheme before them. Affixes are used either to produce "lexeme" from the word they are added or to make the words they are added more functional in the sentence. In this case, determining the morphemes in Turkish means to separate words into morphemes in the form of root and affix.

It is possible to separate the roots and affixes in Turkish into two groups.

1. root morphemes:
    a. nouns
    b. verbs
2. affix morphemes:
    a. affixes to noun
    b. affixes to verb

The morphemes in Turkish can be recognized easily most of the time; however, there are exceptional uses in Turkish as there is in every language.

However, an additional study to obtain the morphemes, the informational values of which were measured, was not made in this study; and the data of the previous study in question were used (Yılmaz, 2018). Because of both the exceptional uses in Turkish and the method followed in the analysis (Yılmaz, 2018: 85-92), in this study, the algorithms used in computers could not be trusted during the separation of words into morhemes. Because computer algorithms can’t distinguish between homogramic words.

In this study, it was created the a corpus of one hundred text fragments selected from proverbs, idioms and poems; from lectures on religion and politics; from news or criticism articles about politics, science, culture, art, economy, sports, weather published in newspapers and magazines; from narratives such as story, novel, memory, fairy tale, anecdote. A total of 34598 morphemes (root and suffix) were classified in 2141 different items by analyzing a collection of text fragments with an average of 26 sentences (2623 sentences in total).

Suffixes are classified in four groups: 1. inflexional suffix to name, 2. derivational suffix to name, 3. inflexional suffix to verb, 4. derivational suffix to verb.

The following principles were followed when analyzing the texts:

1. Proper nouns and exclamations are excluded from the review.

2. The interrogative particle "*mı/mi/mu/mü*", the conjunction "*da/de*" was accepted as the suffix because they fit the sound harmony. The conjunction "*ki*" anda prefixes in words such as "*antipati*" (antipathy), "*bitap*" (exhausted), "*bilfiil*" (de facto), "çün*k*ü" (because), "*gayrimenkul*" (real estate), "*ilelebet*" (forever), "*maalesef*" (unfortunately), "*namert*" (craven), "*ta ki*" (until), "*veya*" (or), "*ya da*" (or), "*yahut*" (or) that entered Turkish from foreign languages, were accepted as roots because they did not conform to their sound harmony.

3. Only categories marked in standard Turkish were analyzed. For example, in modern Turkish there is a morpheme for negativity in verbs while there is no morpheme for positivity. Here it was preferred to observe a suffix of negativity rather than assume that a category exists. However, although not usually marked in standard Turkish, the complementary verb was accepted as a category. Because this category is a category that can be marked in Turkish.

4. Words were analyzed morphologically. A " $+$ " sign was put at the end of the noun roots and a " $-$ " sign at the end of the verb roots. Suffixes were divided into noun suffixes and verb suffixes. A " $+$ " sign was put in front of the suffixes that came to the name, and a " $-$ " sign was placed in front of the suffixes that came to the verb. In addition, the suffixes to the noun and the verb were divided into inflexional and derivational. While doing the analysis, it was stipulated that the roots and suffixes that make up the words should be the morphemes used in standard Turkish.

5. During the use of roots and suffixes in texts, differences in meaning and function were not taken into account. For example, the word "*koca*" is used in texts to mean "husband" or "old man". This difference in meaning was ignored. Similarly, the word "*hayır*" used to mean "benefit" or "negative answer" was also examined in the same item. The roots or suffixes with common origin are collected in the same item. For this reason, etymological dictionaries (or other relevant publications in the literature) were used. The same was done for the suffixes. For example, the suffix "$+$*ca*/+*ce* /+*ça*/+*çe*", which comes to the nouns (Korkmaz, 1960), was examined in a single item, although it is used in different functions in the texts.

6. However, homogramic roots and suffixes were considered as different morphemes. Some of the homogramic roots (such as "*ak*/*ak* $-$", "*eski*/*eski* $-$", "*boya*/*boya* $-$"), which can be both nouns and verbs, are found in walking dictionaries, and some of them can only be identified with the help of etymological dictionaries, since they are no longer used.

7. While doing the analysis, the format changes that occur during use were not taken into account. For example, the ablative is used as four different forms of the same suffix ("$+$ *dan*","$+$ *den*","$+$ *tan*","$+$ *ten*") in standard Turkish. These were collected in the same item. The same method was applied on the roots. Differences in use (as in the words "*kitaptan*" (from the book) and "*kitabı*" ("the book" that is the object of

the sentence)) were not taken into account. Some local or class uses from spoken language into written language were also not taken into account in the analysis. For example, the word "$bilyem$" in the texts was interpreted as "$biliyorum$" (I know), the word "$anicim$" was interpreted as "$anneciğim$" (my mom, mommy) and "$buba$" was interpreted as "$baba$" (father).

8. The affixes deriving verbs from the noun such as "$+lan -/+len -$" in word "$hasta + lan -$", which are often considered as a whole in grammar books, were parsed into morphemes such as "$+la - n -/+le - n -$" (get sick). Because the suffixes that make up these suffixes are the ones used in standard Turkish. Stereotyping in some uses is not important for this study.

9. It is controversial whether the roots of words such as "$birey$" (individual), "$düzey$" (level), "$kuzey$" (north), "$güney$" (south), "$aday$" (candidate), "$yapay$" (artificial), "$olay$" (event), "$uzay$" (space), "$dolayı$" (because of), which were derived after the language revolution in Turkish, are noun or verb. In this regard, using Sertkaya's article (Sertkaya, 2012), the roots of words "$birey$", "$düzey$", "$kuzey$", "$güney$" were accepted as the nouns, the root of the words "$yapay$", "$olay$", "$uzay$", "$dolayı$" is accepted as a verbs, and the word "$aday$" was accepted as a whole noun root. The word "$özgür$" (free), which was also derived after the revolution, was accepted as a compound word as "$öz$" and "$gür$". Although the root of the word "$tasarı$" (design) is found in walking dictionaries as a noun item in the form of "$tasar$", the root of the word was taken as a verb in the form of "$tasar -$" (Gülensoy, 2007: 864).

10. Persian noun phrases such as "$katliam$" (massacre) and "$suikast$" (conspiratorial) and Arabic noun phrases such as "$estağfurullah$" (You're welcome), "$eyvallah$" (thank you), "$inşallah$" (God willing), "$vallahi$" (I swear), "$vesselam$" (that's it), "$maalesef$" (unfortunately), "$bilfiil$" (de facto), which are used as phrases in standard Turkish, were parsed like phrases

FINDINGS (RESULTS)

Coding with Information Theory

It requires to make it clear by a simple fiction why elementary codes are needed and what kind of function they undertake in the coding process in order to be able to understand the evolution of writting. The concepts proposed by the "information theory" have been used for this fiction.

Here two alphabets, each of which can be resembled to a "set", are built. It is aimed at coding the "symbols" in the "source alphabet" (C set) by using "codes". These "elementary codes" are the "elements" of the "set" in the "code alphabet" (S set). The element number of C is much higher compared to that of S and it is expected that this difference will gradually increase in time. Therefore, each element of C can be coded with the use of more than one element of S.

This case can be explained by a simple example:

Example 1

Let each element of the set $C = \{a, b, c, d, e, f, g, h\}$

be c_i in the way $0 < i \leq 8$.

Let each element of the set $S = \{0,1\}$

be s_j,s_k, s_l in the way $0 \leq j < 2$, $0 \leq k < 2$ ve $0 \leq l < 2$.

Since the element number of C is more than that of S, it is clear that more than one element of S should be used if it is wished that the elements of C will be shown with the elements in S.

Since

$s(C) = 8$ and

$s(S) = 2$,

the codes in S must be used three times in order to meet each symbol in C.

By this method, the symbols in C are coded as following by using the codes in S.

"a" $\rightarrow 000$

"b" $\rightarrow 001$

"c" $\rightarrow 010$

"d" $\rightarrow 011$

"e" $\rightarrow 100$

"f" $\rightarrow 101$

"g" $\rightarrow 110$

"h" $\rightarrow 111$

As seen, "$s_k s_l$" unity is needed for each c_i. This cooccurrence is called as "block code" in the informational theory (Abramson, 1963: 46).

As seen above, the "$s_j s_k s_l$" may take same values. That is, the elements in S are "potential block codes", each "repeated permutation" combined of 3 of which can be used in order to code the elements in C.

Therefore, if the element number of S is r and

the element number of C is m,

the condition $r^n \geq m$ should be provided. r^n means the repeated permutation with n of the elements in S.

Since $2^3 \geq 8$, this condition is provided.

In this case, since each unity is composed of three codes, the code, which is as

"$edfbefafhdgefdcefadefgbcgefhefeg$"

and formed of the elements of C, is composed of

$32 \cdot 3 = 96$ codes when coded with these block codes obtained by using the elements in S.

On the other hand, the code number in block codes used to code each symbol will increase because of the fact that the symbol number in the source alphabet (in C) is in the tendency to increase.

This case can be explained by an example:

Example 2

This time, let

$m = 16$ and

$r = 2$.

In this case, the symbols of a source alphabet such as

$C = \{a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p\}$ are coded as:

$"a" \to 0000$	$"e" \to 0100$	$"i" \to 1000$	$"m" \to 1100$
$"b" \to 0001$	$"f" \to 0101$	$"j" \to 1001$	$"n" \to 1101$
$"c" \to 0010$	$"g" \to 0110$	$"k" \to 1010$	$"o" \to 1110$
$"d" \to 0011$	$"h" \to 0111$	$"l" \to 1011$	$"p" \to 1111$

Now, it is necessary to use the block code composed of 4 codes such as "$s_j s_k s_l s_m$" for each c_i.

Then, since the block code which is the equivalent of each symbol is composed of 4 codes, a message again composed of 32 symbols such as

"$pahokahagagakaebabecliamofidefec$" is coded as

$32 \cdot 4 = 128$.

The difference between two coding examples can be seen in the below table:

Table 1. Comparison of "Example 1" and "Example 2"

	Example 1	Example 2
the element (symbol) number of C	8	16
the element (code) number of S	2	2
the code number in every block code	3	4
the number of the potential block code which is not used	0	0
the symbol number in the message	32	32
the code number used to code the message	96	128

As it can be understood, the source alphabet which has more elements increases the code number in block codes and, hence, the code number in the message. The increase of the code number in messages means time, energy and cost increase in the coding process. Therefore, all repeated permutations evolves in a way to slow the code number in messages throughout history (Gemalmaz, 1982: 28)

At a first glance, it can be thought that it is necessary to increase the element number in the code alphabet in order to realize the mentioned slowing down.

This method is used in the below example:

Example 3

Let the element number of S be increased by one in a way to code the elements of C in the "Example 1". It is possible to code the elements of C by a code alphabet of 3 elements such as $S = \{0, 1, 2\}$.

$"a" \to 00$	$"e" \to 11$
$"b" \to 01$	$"f" \to 12$
$"c" \to 02$	$"g" \to 20$
$"d" \to 10$	$"h" \to 21$

As seen, code blocks ($s_j s_k$) composed of twice-repeated permutations of elements in S are sufficient for each c_i.

That is, since

$3^2 \geq 8,$

$r^n \geq m$ condition is provided.

It should be paid attention here to the fact that

$3^2 = 9$ and

$9 - 8 = 1$ twice-repeated permutation (potential block code) is not used. "22" permutation is not needed as a block code.

Now, the message in "Example 1" can be coded by less codes with the block codes obtained by using S composed of 3 elements. Since the block code which is the equivalent of each symbol is composed of 2 elements here, it is possible to code the the message composed of 32 symbols as

$32 \cdot 2 = 64.$

This difference can be seen in the below table:

Table 2. Comparison of "Example 1" and "Example 3"

	Example 1	Example 3
the element (symbol) number of C	8	8
the element (code) number of S	2	3
the code number in every block code	3	2
the number of the potential block code which is not used	0	1
the symbol number in the message	32	32
the code number used to code the message	96	64

The same comparison may again be given by an example:

Example 4

Let the elements of C in "Example 2" be coded by using the elements of

$S = \{0, 1, 2\}$ this time.

$"a" \rightarrow 000$	$"e" \rightarrow 011$	$"i" \rightarrow 022$	$"m" \rightarrow 110$
$"b" \rightarrow 001$	$"f" \rightarrow 012$	$"j" \rightarrow 100$	$"n" \rightarrow 111$
$"c" \rightarrow 002$	$"g" \rightarrow 020$	$"k" \rightarrow 101$	$"o" \rightarrow 112$
$"d" \rightarrow 010$	$"h" \rightarrow 021$	$"l" \rightarrow 102$	$"p" \rightarrow 120$

Since

$m = 16$ and
$r = 3$
here,
$r^n \geq m$ condition can be provided in the case that
$n = 3$ and
$3^3 \geq 16.$

However, since

$3^3 = 27$

this time, it is seen that

$27 - 16 = 11$ repeated permutation of 3 (potential block code) is not used since

$3^3 = 27.$

The use of "121", "122", "200", "201", "202", "210", "211", "212", "220", "221", "222" permutations is unnecessary.

Similarly, less codes should be used the message in "Example 2" with the block codes obtained by using S composed of 3 elements. Since the block code, the equivalent of each symbol, is composed of 2 codes, the code composed of 32 symbols can be coded as

$32 \cdot 3 = 96$ this time.

And this difference can be shown by a table:

Table 3. Comparison of "Example 2" and "Example 4"

	Example 2	Example 4
the element (symbol) number of C	16	16
the element (code) number of S	2	3
the code number in every block code	4	3
the number of the potential block code which is not used	0	11
the symbol number in the message	32	32
the code number used to code the message	128	96

As it can be understood from all these examples and tables, the unpreventable increase in the symbol number of the source alphabet brings the increase in the code number used to code messages together (see "Table 1"). It seems that it works at the beginning to increase the code number of the code alphabet in order to slow down the increase in the code number of messages (in this way, for saving energy and decreasing cost). (See "Table 2" and "Table 3")

However, the symbol number of the source alphabet increases continuously. By this method, the increase in the code number of the code alphabet will be continuous. Yet, the code alphabet should be composed of a limited number of codes and the code number of this alphabet should be as constant as possible. This limitation is designed in a way to remove the heavy load in human mind, which can be caused by the fact that each sign is coded by a separate code as a result of the unpreventable increase in the quantity of signs. Otherwise, the principal difference between the code alphabet and the source alphabet disappears and the codes in the code alphabet lose their characteristics of being articulated in time.

In this case, the possibility should be searched in order to slow down the increase (the mentioned increase can be slowed down but cannot be stopped) in the code number in messages despite the increase in the symbol alphabet without increasing the code number in the code alphabet. Here the information theory has been developed to achieve this purpose. This theory relates the code number of the block code which is the equivalent of a symbol to the "informational value" of the symbol in question.

Then, it should be explained what the informational value of a symbol is and how this is calculated.

Based on the condition $r^n \geq m$,

Hartley said that

$I(c_i) = log_r m$ (Hartley, 1928: 536-540).

Here the code number of the code alphabet is a "logarithm base", the symbol number of the symbol alphabet is "logarithm magnitude".

$I(c_i)$ is the informational value of each symbol in the symbol alphabet.

By Hartley’s measurement, the informational value of each symbol in the above examples is determined respectively as

$log_2 8 = 3\, bit/symbol$in "Example 1"

$log_2 16 = 4\ \ bit/symbol$ in "Example 2"

$log_3 8 = 1{,}8928\ \ trit/symbol$ in "Example 3"

$log_3 16 = 2{,}5237\, trit/symbol$ in "Example 4".

If the informational value of a symbol is an integer, it gives the code number in the block code which is its equivalent. If the informational value of a symbol is an integer, this value is equated to the closest whole number bigger than itself.

In this case, there should be

3 codes each of which has the value of 1 bit in "Example 1"
4 codes each of which has the value of 1 bit in "Example 2"
2 codes each of which has the value of 1 bit in "Example 3"
3 codes each of which has the value of 1 bit in "Example 4"
in the block code, equivalent of each symbol.

As it is seen, the code numbers in the block code which is the equivalent of each symbol according to the informational value determined by Hartley's method are same as the results in "Table 1", "Table 2" and "Table 3". This method was very valuable for sure and an inspiration for future measurements. However, it was impossible to achieve code saving aimed by the information theory and to explain the evolution of repeated permutation by the approach which is the basis of this measurement.

In his article titled "A Mathematical Theory of Communication" where he laid the foundations of the information theory, Shannon developed a much more practical approach by determining the informational values of symbols not according to the symbol number in the symbol alphabet but according to the "frequency" of symbols in messages (Shannon, 1948: 4-8).

This approach can be explained by a simple example:

Example 5

The informational values of the symbols in "Example 1" should be determined this time according to the frequency of symbols in the message by using the elements of the set $S = \{0, 1\}$ (codes in the code alphabet) in the same example.

For this purpose, the frequency of each symbol in the message should be determined first.

In the message composed of 32 symbols in "Example 1" are used as

"a", 2 times;	"e", 8 times;
"b", 2 times;	"f", 8 times;
"c", 2 times;	"g", 4 times;
"d", 4 times;	"h", 2 times.

As the total symbol number in the message will be q,

the frequency of each symbol in the message will be $P(c_i)$ and the element number of the code alphabet is r,

the informational value of each symbol can be determined by

$P(c_i) = \frac{f(c_i)}{q}$ and

$I(c_i) = -log_r P(c_i)$ (Abramson, 1963: 12).

By using this equation the informational value of the symbol "a" is

$$I("a") = -log_r P(a) = -log_2 P\left(\frac{2}{32}\right) = 4\ bit/symbol.$$

By using the same equation, the informational values of all symbols are respectively determined as

$I("a") = 4\ bit/symbol,$

$I("b") = 4\ bit/symbol,$

$I("c") = 4\ bit/symbol,$

$I("d") = 3\ bit/symbol,$

$I("e") = 2\ bit/symbol,$

$I("f") = 2\ bit/symbol,$

$I("g") = 3\ bit/symbol,$

$I("h") = 4\ bit/symbol.$

Since the informational values of all symbols are whole numbers, the informational value of each symbol is accepted as the code number in the block code which is its equivalent. That is, there should be 4 codes in the block codes equivalent of "a", "b", "c", "h" symbols, 3 codes in the block codes equivalent of "d", "g" symbols, and 2 codes in the block codes equivalent of "e" and "f".

If attention is paid, the informational value (and depending on this, the code number in the block code equivalent of symbol) decreases as the frequency of symbol increases. In this way, the code number in messages decreases.

As the element number of the symbol alphabet is m,

the code number in the block code, equivalent of the symbol, is $s(c_i)$,

the frequency of the symbol in the message is $f(c_i)$,

the message in "Example 1" can be coded by

$\sum_{i=1}^{m} f(c_i). s(c_i) = 88$ codes according to these results.

If a comparison is made, there is a cost saving of

$96 - 88 = 8$ codes in the same message by this method.

The factor that shortens the message by 8 codes here is that the frequency of the symbols in the message is taken into consideration. In this way, "the uncertainty of the system" (the entropy of the system) has decreased. Shannon thought the uncertainty of the system as the "average informational value of the symbols" used in a message.

If the informational value of the symbols in the message is represented by H,

then

$H = \sum_{i=1}^{m} P(c_i).I(c_i) = -\sum_{i=}^{m} P(c_i).log_r P(c_i)$ (Shannon, 1948: 10).

If so, in the coding in "Example 5",

$H = 2.75\ bit/symbol.$

When this value is compared with

$H = 3\ bit/symbol$in the coding in "Example 1", it is seen that the uncertainty ratio of the system in "Example 5" is less compared to that of "Example 1" since the frequency of the symbols in "Example 1" in the message is not taken into consideration.

In other words, it is supposed that the frequency of each symbol is equal to each other. This is the case where the uncertainty of the system is highest. That is, for Hartley, $H = log_r\ m$.

The results of codings in "Example 1" and "Example 5" can be compared by a table:

Table 4. Comparison of "Example 1" and "Example 5"

	Example 1.	Example 5.
the element (symbol) number of C	16	16
the element (code) number of S	2	2
the symbol number in the message	32	32
the code number used to code the message	96	88
the average informational value of the symbols	$3\ bit/symbol$	$2{,}75\ bit/symbol$

As it is understood from the table, coding the symbol according to frequencies lessens its uncertainty and, therefore, the code numbers in messages.

As it will be understood from "Example 5", the informational value of the symbol, the frequency of which increases, decreases. The symbols with low informational values need other symbols in order to pass information in the communication process; thus, they become more articulated. In this case, it may be needed to know with which symbol (or symbols) as well as with which frequency the symbol in question is used besides knowing its frequency. That is, the more data about the use of a symbol are, the more the safety of the measurement made for the informational value of a symbol is. For this reason, Shannon measured the informational values of letters and words by determining the 2, 3, 4 times repeated letter permutations of the English letters and 2 times repeated word permutations of the words (Shannon, 1948: 4-8).

It is necessary to objectify this measurement by an example:

Example 6

The informational values of the symbols "e" and "f" have been determined as 2 $bit/symbol$ in the message in "Example 5" and it was mentioned that the block codes meeting both should be composed of 2 codes (see "Example 5"). Now, the informational values of these 2 codes will be measured as "ef".

For this purpose,

$$P("ef") = p("e") \cdot P_{"e"}("f")$$

gives the frequency (probability) of "ef" on condition that the frequency (or probability) of "e" is $P("ef")$ and the frequency (or its conditional probability) is $P_{"e"}("f")$.

Then when

$$P("e") = \frac{8}{32},$$

$$P_{"e"}("f") = \frac{6}{8},$$

$$P("ef") = \frac{8}{32} \cdot \frac{6}{8} = \frac{48}{256}.$$

$P("ef")$ is called as "compound probability" here. Now the informational value of "ef" can be found.

If

$P("ef") = \frac{48}{256}$ and

$r = 2,$

$I("ef")$ is measured as

$$I("ef") = -log_r P("ef") = -log_2 P\left(\frac{48}{256}\right) = 2{,}41504\ bit/symbol.$$

When this value is integrated into the closest integer bigger than itself,

$$s("ef") = 3.$$

When it is considered that

$s("e") = 2,$

$s("f") = 2$ and

$s("e") + s("f") = 2 + 2 = 4$ codes,

it is seen that

$$s("ef") < s("e") + s("f").$$

That is, the block code number which is the equivalent of "ef" decreases by 1 code when the symbols "e" and "f" are coded as they are thought to be symbols.

Informational Value of Morphemes in Turkish

Since it is obligatory to know its frequency in the corpus in order to measure the informational value of any morpheme, morphemes should be separated into 2 sets as root and affix because affix (or affixes) are connected to roots. In a clearer expression, when it is thought that each morpheme is like a ball in a bag (Here the corpus is like a bag), there is no possibility that any ball to be taken from the bag will be an affix. That is, the first ball to be taken from the bag will be a root. If so, a set (C set called as the "symbol alphabet" in the "Theory" part) composed of roots should first be prepared since the system alphabet will be composed of roots.

In this case, the division of the frequency of a root (f(c_i)) into the total root number of those in the corpus ($q = 16921$) gives the frequency of the root ($\mathrm{P(c_i)}$) in question.

Finally, the code alphabet defined in the "Theory" part is thought as $S = \{0, 1\}$ in this part in order to code any symbol in this corpus. That is, the informational values of symbols were measured in terms of "bit" ($r = 2$).

Then, the informational value of any root morpheme can be measured by the data of this study.

Example 7

The informational value of the root morpheme "$y$üz" (meaning "number 100" as well) in the corpus can be measured by using the dataof the study in question (Yılmaz, 2018: 125). In Turkish, "yüz", "homogramic" (Şeka, 2011: 165) morpheme. It means "number 100" and "face".

As in the way that

$$q = 16921,$$

$$f(\text{"}y\text{üz"}) = 67,$$

$$r = 2,$$

$$P(\text{"}y\text{üz"}) = \frac{f(\text{"}y\text{üz"})}{q} = \frac{67}{16921} \text{ and}$$

$$I(\text{"}y\text{üz"}) = -log_r P(\text{"}y\text{üz"}) = -log_2 P\left(\frac{67}{16921}\right) = 7{,}98\ bit/symbol.$$

When this number is equated to the closest integer which is bigger than itself,

$$S(\text{"}y\text{üz"}) = 8.$$

Another measurement can be made by these results in order to make a comparison:

Example 8

Let the informational value of the root "$y\ddot{u}z$" (meaning "face" as well) morpheme which is the homonymic of the root in "Example 7" be measured by the data of the study in question (Yılmaz, 2018: 125).

As in the way that

$$f(\text{"}y\ddot{u}z\text{"}) = 19,$$

$$\mathrm{I}(\text{"}y\ddot{u}z\text{"}) = -log_2 P\left(\frac{19}{16921}\right) = 9{,}798\ bit/symbol.$$

When this result is equated to the closest integer bigger than itself,

$$s(\text{"}y\ddot{u}z\text{"}) = 10.$$

The results in the last two examples are compared in the below table:

Table 5. Comparison of "Example 7" and "Example 8"

	Example 7	Example 8
frequency	67	19
informational value	7,98 $bit/symbol$	9,798 $bit/symbol$
the code number in the block code	8	10

As it is seen, the informational values and the code numbers in the block code which is their equivalent are different since the frequencies of the roots "hundred/face" are homogramic are different.

It is known that the word is composed of the combination of root and affix (or affixes) in Turkish. How is the informational value of the word composed of root and affix (or affixes) measured in this case?

It will be more beneficial to show this by an example:

Example 9

The morphemes of a word should first be determined in order to measure the informational value of the word "$yerlerini$" (their places) chosen from the corpus and to determine the code number in the block code which is the equivalent of the word, depending on this measurement.

The word "$yerlerini$" is analyzed as

"yer", noun root;

"$-ler$", plural affix;

"$-in$", third person singular possession affix (Pronominal "-n" is evaluated within the possession affix here);

"$-i$", loading (or assertion) affix.

It is necessary to calculate the compound probability as in "Example 6" in order to measure the informational value of the word "$yerlerini$" since it is composed of more than one morpheme.

As in the way

$$P("yer") = \frac{73}{16921},$$

$$P_{"yer"}(" - ler") = \frac{9}{73},$$

$$P_{"yer-ler"}(" - in") = \frac{2}{9},$$

$$P_{"yer-ler-in"}(" - i") = \frac{1}{2},$$

$$P("yerlerini") = \frac{73}{16921} \cdot \frac{9}{73} \cdot \frac{2}{9} \cdot \frac{1}{2} = \frac{1314}{22234194}.$$

Now the informational value of the word "$yerlerini$" can be measured out of the calculated compound probability.

$$I("yerlerini") = -log_2 P\left(\frac{1314}{22234194}\right) = 14{,}046\ bit/symbol.$$

When this result is equated to the closest integer bigger than itself

$$s("yerlerini") = 15.$$

If it is paid attention, the frequency of the word "yer" root is determined first while calculating the compound probability here; then, the frequencies of the related affixes (their "conditional probabilities") are determined out of the frequency of "$-ler$" affix in meeting "yer" root in the corpus and of the frequency of "$-i$" affix meeting the word "$yerlerin$" in the corpus. That is, conditional probabilities are calculated out of the frequency together with the related affix (or word) in the culture, corpus in question.

This approach related to conditional probabilities is different from the previous study. In the previous study, conditional probabilities were calculated out of the frequency together with the related root (or word) in the culture, corpus. The fact that the frequency in the culture, corpus in question attaches to the related affix (or word) according to Turkish grammar was calculated out of the total frequency ratio of all possible affixes in the corpus (for application see "4th Example"; Yılmaz, 2018: 64-68).

As it will be understood, in the approach posed in the calculation of conditional probability in the previous study, the basic assumption of the information theory (Sedor, 2015) was sacrificed in order to obtain results closer to reality by considering the insufficiency of the corpus. In this study, the principles of the information theory was based on by moving from the assumption in question. However, it is suitable to focus on the method rather than the obtained results since the corpus is not large enough to put forward the frequencies of Turkish morphemes.

Turkish "predicate attraction" can be examined by an example:

Example 10

The word "$geldim$" (I came) seen in the corpus was analyzed in the previous study as

"gel"; verb root;

"$-di$", past tense verbal adjective affix;

"$-ø$", complementary verb (This category is sometimes coded in Turkish. Predicate attraction cannot be made without "complementary verb" in Turkish.);

"$-m$", 1st person singular pronoun.

For the conditional probability, the possibility of the root and that of the affixes attached to the root (or word) are:

As in the way that

$P("gel") = \frac{161}{16921},$

$P_{"gel"}("-di") = \frac{25}{161},$

$P_{"gel-di"}("-ø") = \frac{25}{25},$

$P_{"geldi-ø"}("-m") = \frac{6}{25}$ and

$P("geldim") = \frac{161}{16921} \cdot \frac{25}{161} \cdot \frac{25}{25} \cdot \frac{6}{25} = \frac{24150}{68107025}.$

Now, the informational value of the word "$geldim$" can be measured based on the calculated compound probability.

$$I("geldim") = -log_2 P\left(\frac{24150}{68107025}\right) = 11{,}461\ bit/symbol.$$

When this result is equated to the closest integer bigger than itself,

then

$s("geldim") = 12.$

The point to be paid attention here is that the conditional probability of the necessary complementary verb ($-ø$) is equal to 1. Although the complementary verb in Turkish is sometimes coded (verb "$-imek$"), it generally disappears. Therefore, the effect of the complementary verb in Turkish predicate attraction on the informational value of the word (and on the code number in the block code which is the equivalent of the word) is almost none in terms of the information theory.

This ineffectivity is valid not only for affixes but for roots as well. This can be seen in the below example:

Example 11

The verb "$tebrik\ et$" (to congratulate) is composed of two morphemes (the noun "$tebrik$" and the complementary verb "et") in Turkish. This means that the morphemes in question, which form the verb in question, can be articulated with different words.

However, when the corpus in question is checked, it is seen that

$P(\text{"}tebrik\text{"}) = \frac{3}{16921},$

$P_{\text{"}tebrik\text{"}}(\text{"}et\text{"}) = \frac{3}{3}.$

That is, the morpheme "et" is used everywhere the morpheme "$tebrik$" is used.

In this case,

since it will be

$P(\text{"}tebrik\text{"}) = P(\text{"}tebrik\ et\text{"}) = \frac{3}{16921},$

there will be

$I(\text{"}tebrik\text{"}) = I(\text{"}tebrik\ et\text{"}) = 12{,}461\ bit/symbol.$

In the same way, there will be

$s(\text{"}tebrik\text{"}) = s(\text{"}tebrik\ et\text{"}) = 13.$

No matter the fact that the expression "$tebrik\ et$" is analyzed as 2 morphemes ("$tebrik$" and "et") according to the Turkish dictionary, its usage condition in the corpus shows that there are not 2 but 1 morpheme here.

As it is seen, the information theory makes it possible to determine new morphemes composed of more than one morpheme’s becoming stereotyped on condition that reliable data are obtained on a large enough corpus.

The Relationship between Entropy and the Alphabet

A calculation developed by Sahannon for the average informational value of symbols (entropy) in addition to the informational value of each of them was mentioned before. Depending on this calculation,

$H_K = \sum_{i=1}^{m} P(c_i).I(c_i) = 9{,}1\ bit/symbol$

as in the way that

the number of different root morphemes in the corpus is $m = 1951$,

the frequency of each root morpheme in the corpus is $f(c_i)$,

the total root number used in the corpus is $q = 16921$, the frequency of the use of each root morpheme in the corpus is $P(c_i) = \frac{f\ (c_i)}{16921}$,

the informational value of each root in the corpus is $I(c_i) = -log_r P(c_i)$,

the code number in the code alphabet is $r = 2$,

and the average informational value of the root morphemes used in the corpus is H_K.

When this result is equated to the closest integer bigger than itself,

$s(H_K) = 10$ (Yılmaz, 2018: 76).

The average code number ($s(H_K)$) in the block codes which are equivalent of the roots in the corpus, obtained out of the average informational value of the roots (H_K) in the corpus, is very important for the repeated permutation of Turkish since this number will be used in order to establish the relationship between the letter number in the Turkish alphabet and the code number necessary to code the root morphemes.

According to this result, a table such as the following can be prepared related to the relationship in question:

Table 6. Comparison of Five Different Options

	1th option	2nd option	3rd option	4th option	5th option
the letter number in the alphabet	2	4	8	16	32
the informational value of each letter as (bit)	1	2	3	4	5
the letter number necessary to code each root morpheme	10	5	4	3	2

As it will be understood from "Table 6", "the letter number necessary to code each root morpheme" decreases as "the letter number in the alphabet" increases because "the informational value of each letter" increases as "the letter number in the alphabet" increases.

The average informational values of the roots in the corpus (H_K), $r = 3$ can be calculated out of (that is as "$trit$").

In this case,

$$H_K = \sum_{i=1}^{m} P(c_i).I(c_i) = 5{,}741\ trit/symbol.$$

When this result is equated to the closest integer bigger than itself,

$s(H_K) = 6$ (Yılmaz 2018: 77).

According to this result, the table is as following:

Table 7. Comparison of Three Different Options

	1th option	2nd option	3rd option
the letter number in the alphabet	3	9	27
the informational value of each letter as (trit)	1	2	3
the letter number necessary to code each root morpheme	6	3	2

The relationships among "the letter number in the alphabet", "the informational value of each letter" and "the letter number necessary to code each root morpheme" in "Table 6" are valid for "Table 7" as well.

CONCLUSION and DISCUSSION

In this study, a different approach based on information theory is presented about an activity that aims to divide language signs (morphemes) into their elements, called "second articulation". Since the function of the morpheme to be analyzed is not taken into account, it is impossible to find a functional feature in a sound piece obtained as a result of an analysis made according to the articulation of this morpheme. In such an analysis, the paradigm is reached independently of the system. Because sound parts are classified according to their articulation, a category can only be created with sound parts whose articulation properties are close to each other. Since sound parts selected from such a category cannot be elements, morphemes cannot be created with them.

The focus should be on the use of morphemes in the language system in order to determine the information value of the morphemes of language, which is itself a system, considering that the system is a whole that is reconstructed and evolved in this way each time. Now the second articulation with the concept of "phoneme" is unthinkable, at least initially. The codes that serve to distinguish the morphemes from each other can only be determined by considering their meaning. In this case, there is a need for a concrete definition of the concept of "meaning". This definition was in a theory put forward by Shannon. He defined the information that is used to distinguish one element in a set as the meaning of that element. His approach presented a method for determining how many codes an element in the set can be encoded with, based on the usage probability of that element.

Based on Shannon's approach, in this study, each morpheme is considered as an element of a set of morphemes. In this case, a set of morphemes was needed. For this reason, two sets (set of roots and suffixes) were formed from these morphemes by dividing them into corpus morphemes, which consisted of a hundred text fragments that were published in the last decade and were close to spoken language. The frequency of each morpheme in these sets was considered as its probability. From this probability, the information value of each morpheme was determined using Shannon's logarithmic measurement method. This measurement was easier in the roots. The ratio of the usage frequency of each root to the usage frequency of all roots provided enough data. However, with the use of suffixes, it became difficult to measure. The suffix that came to each root had to be proportioned to the frequency of use of the elements of a set of all suffixes that came to that root. In other words, it was necessary to create a different set for the suffixes each time. The probability of the words could be determined by multiplying the obtained ratios (ratios for roots and suffixes) with each other. The informational value of words could only be measured by these probabilities.

In addition to the information value of each word, the entropy of the set of roots was also determined using the equation that Shannon obtained from the average of the information values of the elements of the set. The average information value (enropy) was used to determine how many codes should be in the alphabet to represent the words and what the information value of each code was in relation to the number of codes in the alphabet.

At this stage of the study, one of the two options should be selected. The first method is to create a set of suffixes that come to the root and determine the usage frequency of any suffix based on the usage frequency of the suffixes in this set. In this method, it is assumed that the corpus covers all texts produced and to be produced in the language. In the other method, based on the information in grammar books, a set of suffixes suitable to come to a root is formed. In this method, it is accepted that the corpus used does not cover all the texts of the language. It is an error to determine the morphemes in the corpus according to the Turkish grammar theoretically because a morpheme should be a part that can be articulated. These parts should be determined not by the grammar rules but by an algorithm which uses statistical methods out of the frequency. (The dilemma in "Example 11" results from this point). While the first method is theoretically more consistent, the second method is more practical.

Besides, it is a theoretical error to collect morphemes in two sets as "root morphemes" and "affix morphemes". While root and affix morphemes are used together in the corpus, the informational values of morphemes are measured by depending on a set where roots are collected (see "1th Table (The Table of Roots)"; Yılmaz, 2018: 93-126). Since affix morphemes cannot be used independent of root morphemes (that is, since there is no probability that the first ball is an affix morpheme to be taken from the bag if morphemes are thought as balls in a bag), such a way was followed. However, the informational values were found to be less than they should be because the frequency of root morphemes is more.

The fact that affix morphemes are dependent on root morpheme makes it impossible to measure the frequency of the permutations with $2, 3, 4$, etc. of the morphemes in the corpus and the informational values based on these frequencies because the combination of "affix–root" (permutation) is impossible while the combination of "root–affix" (permutation) is possible since Turkish is a head-final language. On the other hand, approaching the real informational values of symbols requires the determination of the frequencies of the permutations composed of more symbols. Shannon determined the letter permutations with $2, 3, 4$ of English words for this reason (Shannon, 1948: 4-8).

In these types of studies, working with large database ensures that the results are more reliable. Because, according to the "Law of Large Numbers", the number of experiments should be increased as much as possible in order to make more reliable predictions about the consequences of an event. In this study, if the number of texts (and therefore the number of sentences) were greater, the probability of morphemes would be closer to the actual use of the language. However, since computer algorithms were not used in text analysis, a limited number of analyzes could be made.

In fact, this work contains many innovations in terms of methods and results, for that matter, is not the first in Turkey. Gemalmaz has an important study titled "Standart Türkiye Türkçesinin Formanlarının Enformatif Değerleri ve Bu Değerlerin İhtiyaç Hâlinde Bu Dilin Gelişimine Muhtemel Etkileri" (Informational Values of Standard Turkish Forms and the Possible Effects of These Values on the Development of This Language When Needed), in which he evaluates the morphemes of Turkish in terms of their informational value. However, these studies are not sufficiently known by the linguist and IT specialists in Turkey. It must be admitted that since Gemalmaz's work was completed, there was not much progress in this matter. Because natural language processing studies (nlp) that will decompose Turkish texts into their morphemes are not at the desired stage yet. Without the help of such algorithms, it is not possible to obtain data to be prepared manually from a large collection. For this reason, such studies should be criticized not for their results, but for their method.

On the other hand, studies were carried out to measure the informational value of Turkish morphemes in order to compress Turkish texts in text processing studies in the field of informatics. The doctoral dissertation entitled "Türkçenin Biçimbilim Yapısına Dayalı Bir Metin Sıkıştırma Sistemi" (The Text Compression System Based on the Morphology of Turkish), prepared by Diri, is remarkable in terms of measuring the informational value of Turkish morphemes based on their frequency of use in texts. However, as the name suggests, since the aim of this study is to develop a digital text compression method that saves lossless data, it requires special effort to interpret the results of such studies linguistically.

RECOMMENDATIONS

On the other hand, the studies which will be made in order to measure the information value with a larger corpus will give results closer to reality. The measurements which have been made will give an idea about the code numbers necessary to code each morphemes. In this way, the change of each morpheme in the coding process will be explainedby relating morphemes to its frequency. In addition, the information value of these morphemes can be measured more closely with an algorithm that can determine the usage frequency of the 2, 3 and 4 combinations of Turkish morphemes with a larger corpus. Thus, the change of each morpheme in the coding process can be explained by correlating it with its usage frequency. As it is understood, these studies necessitate the cooperation of different disciplines.

ETHICAL TEXT

In this article, journal writing rules, publishing principles, research and publishing ethics rules, journal ethics rules are followed. Responsibility belongs to the author for any violations related to the article.

Autors's Contribution Rate: The contribution rate of Çağlayan Yılmaz, the only author of the article, to the article is 100%.

REFERENCES

Abramson, N. (1953). *Information Theory and Coding*. McGraww-Hill Book Company Press.

Barthes, R. (1979). *Göstergebilimin İlkeleri*. (B. Vardar and M. Rifat, Trans.). Ministry of Culture Press.

Diri, B. (1999). Türkçenin Biçimbilim Yapısına Dayalı Bir Metin Sıkıştırma Sistemi [Unpublished doctoral dissetation]. Yildiz Technical University.

Faulmann, C. (2018). *Yazı Kitabı*. (I. Arda, Trans.). Türkiye İş Bankası Cultural Press.

Gemalmaz, E. (1982). *Standart Türkiye Türkçesinin Formanlarının Enformatif Değerleri ve Bu Değerlerin İhtiyaç Hâlinde Bu Dilin Gelişimine Muhtemel Etkileri*. http://efrasiyap.tripod.com/yazilar/STT1.pdf

Gülensoy, T. (2007). *Türkiye Türkçesindeki Türkçe Sözcüklerin Köken Bilgisi Sözlüğü*, Türk Dil Kurumu Press.

Hartley, R. V. L. (1928). "Transmission of information." *The Bell System Technical Journal*, 7, 535-563. http://telecomlaw.ru/young_res/trans_inform_1928.pdf

Ifrah, G. (1998). *Rakamların Evrensel Tarihi*. (K. Dinçer, Trans.). TUBİTAK Press.

İmer, K., Kocaman, A. & Özsoy A. (2011). *Dilbilim Sözlüğü*. Boğaziçi University Press.

Korkmaz, Z. (1960). "Türk Dilinde +ça Eki ve Bu Ek ile Yapılan İsim Teşkilleri Üzerine Bir Deneme." *Ankara Üniversitesi Dil ve Tarih Coğrafya Fakültesi Dergisi*, 17, 275-358.

Lévi-Strauss, C. (2013). *Mit ve Anlam*. (G. Y. Demir, Trans.). Ithaki Press.

Martinet, A. (1998). *İşlevsel Genel Dilbilim.* (B. Vardar, Trans.). Multilingual Press.

Piaget, J. (1982). *Yapısalcılık*. (F. Akadlı, Trans.). Dost Kitabevi Press.

Saussure, F. (1998). *Genel Dilbilim Dersleri.* (B. Vardar, Trans.). Multilingual Press.

Sedor, K. (2015). *The Law of Large Numbers and its Applications*. Lakehead Universty Press.

Sertkaya, O. F. (2012). "Etimoloji Nedir - Ne Değildir ve İsimden İsim Yapan +ay/+ey Eki Üzerine." *Ankara Üniversitesi Dil ve Tarih Coğrafya Fakültesi Türkoloji Dergisi*, 19, 43-72.

Shannon, C. E. (1948). "A Mathematical Theory of Communication." *The Bell System Technical Journal*, 27, 379-473. http://people.math.harvard.edu/~ctm/home/text/others/shannon/entropy/entropy.pdf

Şeka, Y. V. (2011). "Türkiye Türkçesinde Çokanlamlılık ve Eşgösterenlilik." (R. Adzhumerova ve E. Atmaca, Trans.). *The Journal of Language Researches*, 9, 165-172.

Yılmaz, Ç. (2018). *Türkçede Anlam Birimlerinin Bilgi Kuramı Temelinde İşaretlenmesi*. Gazi Kitabevi Press.

https://github.com/COMU/zemberek-extension#giriş