Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2013) 21: 804 – 811

doi:10.3906/elk-1110-11

A new Morse code scheme optimized according to the statistical properties of Turkish

Emrah ÇİÇEK, Asım Egemen YILMAZ*
Department of Electronical-Electronics Engineering, Faculty of Engineering, Ankara University,
06100 Tandoğan, Ankara, Turkey

Received: 08.10.2011 • **Accepted:** 17.02.2012 • **Published Online:** 03.05.2013 • **Printed:** 27.05.2013

Abstract: Morse code has been in use for more than 180 years, even though its currently known form is slightly different than the form defined by Morse and Vail. The code book constructed by Vail was optimized according to the statistical properties of English. In this study, we propose a new code book optimized for Turkish and demonstrate that it is information-theoretically possible to achieve about a 10% improvement throughout the coding of Turkish texts by means of our proposal. The outcomes of this might serve as a basis for potential (academic and/or applied) Turkish language-specific lossless data compression studies.

Key words: Information theory, source encoding, Morse code, language statistics, data compression

1. Introduction

Morse code is a method of transmitting textual information as a series of on–off signals (e.g. tones, lights, clicks), which can be directly interpreted by a skilled listener or observer. This listener or observer might be either a human individual with the knowledge of the code book, or special equipment in which the code book and its encoding/decoding rules are embedded. The code book of the International Morse Code includes the Roman alphabet, the Arabic numerals, and a small set of punctuation marks. According to the code book, a character stream is converted to procedural signals, which are nothing but standardized sequences of short and long "dots" (vocalized as "di"s or "dit"s), and "dashes" (vocalized as "dah"s).

Beginning in 1836, in the United States, an electrical telegraph system was developed by Samuel F. B. Morse, Joseph Henry, and Alfred L. Vail. This system was based on the transmission of electric current pulses along the wires and the control of an electromagnet located at the receiving end. Even though Morse had only planned to transmit numerals in his earliest code, Vail later expanded it in order to include letters and special characters, which yielded a more general usage.

In order to be able to transmit English text with a dot–dash sequence of minimum length, Vail constructed his code book after determining the frequency of the use of the letters in the English language from a local newspaper in Morristown [1]. Vail's effort can be considered as one of the earliest examples of "source encoding", in which redundant data are eliminated from the transmitted information for the economical usage of the resources (i.e. bandwidth in the channel) [2].

Vail constructed the code book by considering the 26 letters in English, and he assigned dot–dash symbols to each of the letters according to its occurrence rate in meaningful English texts. Since many other natural

*Correspondence: aeyilmaz@eng.ankara.edu.tr

languages use more than the 26 Roman letters, extensions to the Morse alphabet were later made for those languages.

In Turkey, the telegraph system has been in use since 1855. For this purpose, the International Morse Code has been used. As stated before, the code book construction of the International Morse Code is based on the language statistics of English; namely, the International Morse Code might not be providing efficient source encoding for the transmission of Turkish texts. Departing from this fact, we try to see what would have happened if a code book optimized for Turkish language characteristics and statistics had been constructed.

Certainly, it can be argued that especially after the development of more advanced data communication means, the telegraph system lost its importance and became almost obsolete in the last decade. On the other hand, studies about the Morse Code (and its performance in other languages) are still valuable from an information-theoretical point of view, since the Morse Code (and its underlying philosophy) constitutes a basis for simple entropy coding, which is considered as the final step of "lossless data compression" applications. The data compression schemes, which are known as Shannon–Fano coding [3,4] and Huffmann coding [5], were developed with inspiration from Vail's effort. Therefore, it is our assumption that the outcome of our study would not be quite meaningful and helpful for potential usage in any telegraph system; rather, it might serve as a basis for potential (academic and/or applied) Turkish language-specific lossless data compression studies.

The organization of this paper is as follows: after this introductory section, we first describe our corpus and summarize our findings about the statistical properties of Turkish, and then we construct our own Morse code book in Section 2. In Section 3, we try to come up with some figures of merit regarding the effectiveness of our efforts. Section 4 will include the concluding remarks of this study.

2. Construction of the Turkish Morse code book

2.1. Language statistics of Turkish

In order to have a code book optimized for Turkish, the statistics of Turkish should be considered first. For this purpose, we have extracted the letter occurrence rates (i.e. monogram statistics) in Turkish by means of a corpus seen in Table 1, which consists of literary/nonliterary and technical/nontechnical actual meaningful Turkish texts.

Table 1. Corpus used for the extraction of the Turkish language statistics.

Title of the text/book	Author	Genre
Kar	Orhan Pamuk	Literary (novel)
Markheim	Robert Louis Stevenson (translated to Turkish by Handan Balkara)	Literary (story)
Besleme	Anton Chekhov (translated to Turkish by Ergin Altay)	Literary (stories)
Köylüler (Mujikler)	Anton Chekhov (translated to Turkish by Zeki Baştımar)	Literary (stories)
Various daily articles	Can Dündar	Nonliterary (newspaper articles)
Various daily articles	Melike Karakartal	Nonliterary (newspaper articles)
Various daily articles	Derya Sazak	Nonliterary (newspaper articles)
Various technical papers about data mining	Various authors	Technical (paper)

The corpus consists of more than 3×10^6 letters (excluding the space characters, punctuation marks, and any other special characters). The n-gram and syllable statistics extracted from this corpus demonstrate good

agreement with previous studies [6–8] about the n-gram and syllable statistics of Turkish. This indicates that the corpus is statistically reliable, and the extracted monogram statistics can safely be used. More information about the content of this corpus as well as more analysis results (i.e. n-gram and syllable statistics with deeper and broader classifications) can be found in [9] and [10].

Table 2 lists the occurrence rates of each letter in meaningful Turkish texts (i.e. the results of the monogram statistics analysis), which is quite critical for the construction of a new Morse Code optimized for Turkish.

Table 2. Occurrence rates of the letters in Turkish.

Order	Letter	Frequency (%)	Order	Letter	Frequency (%)	Order	Letter	Frequency (%)
1	a	11.46	11	m	3.51	21	h	1.11
2	i	9.32	12	y	3.32	22	ğ	1.047
3	e	9.07	13	s	3.15	23	ç	1.046
4	n	7.42	14	u	3.14	24	v	1.01
5	r	7.04	15	b	2.67	25	c	0.92
6	l	6.40	16	o	2.58	26	p	0.87
7	k	4.65	17	ü	1.92	27	ö	0.77
8	d	4.60	18	ş	1.53	28	f	0.49
9	ı	4.56	19	z	1.50	29	j	0.05
10	t	3.60	20	g	1.15			

2.2. Observations about the construction of the International Morse Code

The International Morse Code can be defined by means of the dichotomic table seen in Figure 1. The dichotomic table constitutes not only an elegant and compact means for illustration/enlisting of the code book, but also a good guideline for the decoding procedure.

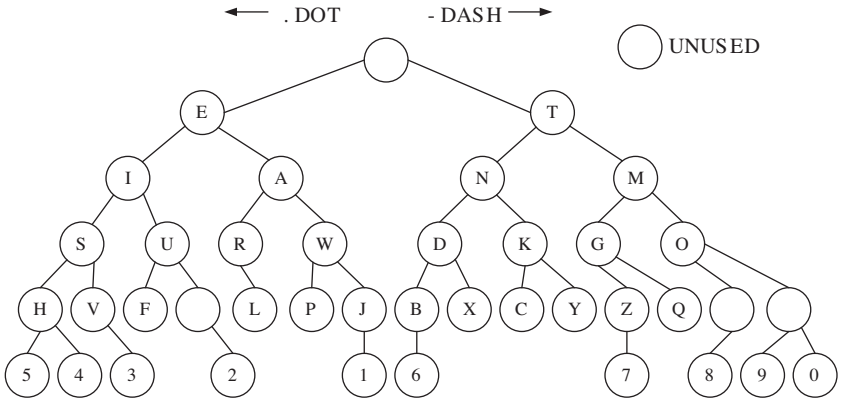

Figure 1. Dichotomic search table of the International Morse Code (adapted from [11]).

According to the dichotomic table, the interpreter/listener branches to the left in case he receives a dot and to the right for a dash, until the transmission of a character/letter is finished. This methodology was developed in order to ease the design of electromechanical Morse Code-interpreting machinery. As an example, if the transmitted character is "Y", the sequence "– . – –" is sent. The receiver starts to track the dichotomic table from the "unused" state seen at the top. After the first "–" is received, the pointer of the decoder goes

down to the right to "T". Then, after the "." is received, the pointer of the decoder goes down to the left to "N". Then, after the "–" is received, the pointer of the decoder goes down to the right to "K". Finally, after the "–" is received, the pointer of the decoder goes down to the right to "Y". This means that the sequence "– . – –" is interpreted correctly as "Y" by definition.

The main idea for the construction of the International Morse Code dichotomic table can be summarized as follows:

- More frequent letters in English have been placed in the upper rows in the dichotomic table during the design. This yields shorter dot–dash representations for more frequent letters and, eventually, better source encoding. In particular, the most frequent letter in English, "E", gets the left position in the upper row. The other letters are sorted according to their frequencies.
- The left–right positioning is based on the following idea: if more frequent letters are placed in the left positions of the branches for some particular row, for the next row, more frequent letters are placed in the right positions of the branches. This is done to achieve a fair transmission duration distribution for the whole alphabet, since the transmission duration for a dash is much more compared to that of a dot.

2.3. Turkish Morse Code

Following a similar philosophy summarized in the previous subsection, we developed the Turkish Morse Code dichotomic table seen in Figure 2.

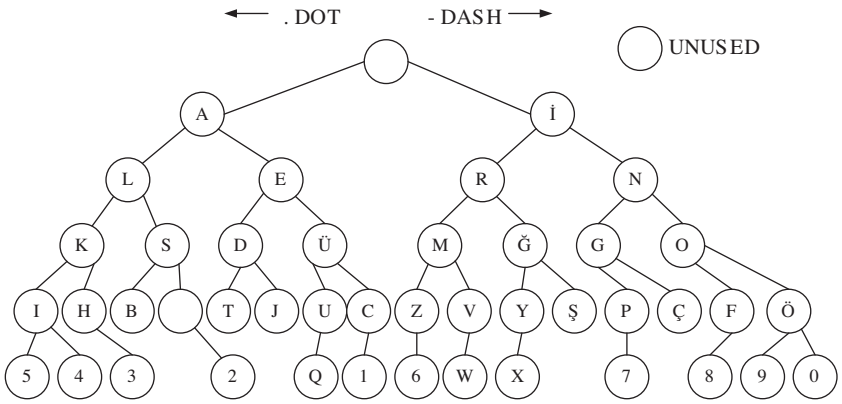

Figure 2. Dichotomic search table of the proposed Turkish Morse Code.

During the construction of the Turkish Morse Code, similar strategies for positioning the letters were followed. There are 2 main objectives during the construction:

1. More frequent letters are positioned in the higher rows in order to represent them with a minimum number of symbols. This action minimizes the total number of symbols required for the representation of an ordinary text in a particular language.
2. For left–right positioning in each branch, it is desired to fill the left branches earlier than the right branches, since the left branch corresponds to a "dot", for which the transmission duration is shorter. This action minimizes the transmission duration of an ordinary text in a particular language via the conventional telegraph system. Certainly, this was quite critical in the days of telegraph communications, but obviously this feature has no particular importance or impact in our study. Nevertheless, for the sake of generality, we followed the same guidelines while constructing our own dichotomic table.

Since "A" and "İ" are the most frequent letters, as mentioned in Table 2, they are positioned in the first row. "A" gets the left position in the branch since it is more frequent than "İ". Next, the letters "E", "N", "R", and "L" are placed in the second row. This time, "E" and "N" get the right positions in each branch, since they have greater frequencies compared to "R" and "L". The other rows are constructed in a similar manner with the construction of the International Morse Code: according to the 2nd objective mentioned above, we place "K" and "D" in the first 2 left branches in the 3rd row, and the succeeding 2 letters "I" and "T" are placed in the left branches in the 4th row, which are connected to "K" and "D", respectively. In this manner, the left branches are filled earlier than the right branches. Next, the available right branches in the 3rd and the 4th row of the dichotomic table are filled. Here, exactly the same procedure proposed by Vail is followed. For example, in our own dichotomic table, we put the 16th most frequent letter of Turkish in the place where Vail positioned the 16th most frequent letter of English in his dichotomic table.

In order to have full coverage for the English alphabet, the letters "Q", "W", and "Z" are also placed after the completion of the placement of the 29 letters in the Turkish alphabet.

3. Results

In this section, in order to demonstrate the effectiveness and the benefits of the Turkish Morse Code, we evaluate and compare the performance of 3 different Morse coding schemes applied to a lengthy (having 1,372,719 characters) but meaningful text written in Turkish. These schemes are described in the upcoming subsections.

3.1. Evaluated schemes

3.1.1. Scheme 1: application of the International Morse Code

In this scheme, the Turkish text is coded by the International Morse Code seen in Figure 1. Since the International Morse Code code book in its pure form does not include Turkish characters, the Turkish characters inside the text are converted to similar English characters. The conversions are defined and performed as follows:

$$\text{ç} \rightarrow \text{c} \quad \text{ğ} \rightarrow \text{g} \quad \text{ı} \rightarrow \text{i} \quad \text{ö} \rightarrow \text{o} \quad \text{ş} \rightarrow \text{s} \quad \text{ü} \rightarrow \text{u}$$

Such conversions will inevitably cause data loss, and sometimes even yield ambiguities at the decoder (receiver) side. For example, with this scheme, the words "aşı" (vaccine) and "asi" (rioter) would be encoded and transmitted in the same way, as "asi". The words "atık" (waste) and "atik" (agile) would be encoded and transmitted in the same way, as "atik".

It should be noted that this scheme was being used (mainly for economic purposes) in the days when the telegraph system was still in use and was the most rapid method for data communications. That is why we find it noteworthy to include this scheme and use it as a reference for comparison in our analyses.

3.1.2. Scheme 2: application of the extended International Morse Code

The International Morse Code has been extended in order to support non-English languages. However, these extensions yielded the assignment of lengthy sequences for local letters, which certainly degrades the source encoding performance. In this scheme, the Turkish text is coded by means of the extended International Morse Code, where the symbols corresponding to Turkish characters are listed in Table 3.

Table 3. Representation of the Turkish characters in the extended International Morse Code [11].

Letter	Representation	Letter	Representation	Letter	Representation
ç	-.-..	ı	.-..-	ş	.–..
ğ̆	–.-.	ö	—.	ü	..–

3.1.3. Scheme 3: application of the proposed Turkish Morse Code

In this scheme, the Turkish text is directly coded by the proposed Turkish Morse Code, which is defined by the dichotomic search table seen in Figure 2.

3.2. Performance metric definition

In order to evaluate the effectiveness of each scheme in a quantitative manner, a performance metric will be defined. This metric will indicate how successfully the source encoding activity is performed.

The symbol-to-character ratio (SCR) might be considered as a performance metric for evaluation of the success of each scheme. It can be defined as follows:

$$SCR = \frac{n_s}{n_c}, \tag{1}$$

where n_c is the number of characters in the text to be encoded and n_s is the number of symbols (i.e. total number of dashes and dots) in the encoded stream.

On the other hand, this quantity with its current form does not consider whether there exists any information loss in a coding scheme. A fair performance metric should also consider and penalize the information loss rate. For this purpose, an effective SCR (denoted by SCR_{eff}) is defined:

$$SCR_{eff} = \frac{n_s}{n_c}(1-\rho)^{-1}, \tag{2}$$

where the information loss rate ρ is defined as follows:

$$\rho = \frac{n_x}{n_c}. \tag{3}$$

In Eq. (3), n_x is the number of characters converted to other characters existing in the code book before encoding. In the case of no information losses, the effective SCR is equal to the SCR. For cases where information losses exist, as n_x increases, $(1-\rho)^{-1}$ becomes a more dominating factor for the effective SCR. A coding scheme should have a minimal effective SCR, since a minimal effective SCR means nothing but a better "source encoding" performance. Hence, the effective SCR definition seems to construct a good performance metric for the coding schemes.

3.3. Comparison of the schemes

Table 4 lists the numerical success of each scheme for the same lengthy Turkish text. Scheme 1 yields an SCR of 2.166. On the other hand, it causes about 10% information loss due to the Turkish character conversions. Hence, the effective SCR for this scheme is evaluated as 2.407. Another remark should be made at this point: our experiments showed that for an English text of same length (i.e. 1,372,719 characters), this scheme yielded 2,921,340 symbols. Since there is no character conversion and information loss for this case, this corresponds

to an effective *SCR* of 2.128. This means that even though the standard International Morse Code is quite effective for English, it actually constitutes an expensive choice for Turkish.

Table 4. Performance comparison of the 3 different coding schemes.

Scheme	Number of characters in the text (n_c)	Number of symbols in the coded stream (n_s)	Symbol-to-character ratio ($SCR = n_s/n_c$)	Number of converted characters (n_x)	Information loss rate ($\rho = n_x/n_c$)	Effective symbol-to-character ratio (SCR_{eff})
1	1,372,719	2,973,171	2.166	137,573	0.1002	2.407
2	1,372,719	3,289,934	2.397	0	0	2.397
3	1,372,719	2,962,249	2.158	0	0	2.158

Scheme 2 yields an *SCR* of 2.397. This was expected since the number of symbols assigned for the nonstandard characters in the extended International Morse Code is quite high (4 or 5, as seen in Table 3). However, it should be mentioned that by sacrificing the *SCR*, information loss is prevented in this scheme.

Scheme 3, which implements the Turkish Morse Code proposed by us, yields the best effective *SCR* value, since it is designed according to the statistical properties of Turkish. In summary, when the effective *SCR* values are considered:

- Scheme 3 provides a 10.35% performance improvement compared to Scheme 1 (since $|2.407 - 2.158| / 2.407 = 0.1035$), and
- Scheme 3 provides a 9.97% performance improvement compared to Scheme 2 (since $|2.397 - 2.158| / 2.397 = 0.0997$).

As mentioned before, the standard International Morse Code yields an effective *SCR* of 2.128 for English texts, whereas our Turkish Morse Code yields a slightly higher effective *SCR* of 2.158 for Turkish texts. The 1.4% difference between these values is acceptable, and its reason can be defended as follows: the standard International Morse Code focuses on the representation of 26 letters, whereas, for the Turkish Morse Code, this number is 29. Hence, it can be concluded that the defined code book is quite successful in terms of source encoding.

4. Concluding remarks

In this study, we tried to visualize what would have happened if the Morse Code had been designed specifically for Turkish (instead of English). Our results showed that the usage of the International Morse Code (either the standard or the extended) is information-theoretically ineffective and expensive in Turkish telegraph systems. Our code book yields about a 10% improvement in terms of the source encoding compared to the International Morse Code.

As stated before, especially in the last decade, the telegraph system has lost its importance and popularity. Hence, the major contributions made by this study would fall into the area of information theory, particularly language-specific "lossless data compression".

Throughout the performance analyses, we have focused on the encoding phase, but have not explicitly considered the decoding phase. This is because the code book's layout is the main and only factor determining the data compression performance. The decoding operation is standard, straightforward, and independent of

the code book's layout. For the decoding operation at the receiver side, it is sufficient to implement (either as hardware or software) a pointer tracing the dichotomic table according to the incoming symbols ("–" or ".") and capturing the relevant letter.

Another important aspect of this study was the identification of the difference between the "lossless" and "lossy" data compression schemes. The *SCR* metric defined in Eq. (1) considers only the number of characters in the text to be encoded and the number of symbols in the encoded stream; it does not care about the data losses. In order to handle the information losses, an alternative metric (effective *SCR*, or SCR_{eff}) penalizing the information loss rate is defined. This metric can safely be used for the fair cross-comparison of the lossless and lossy data compression methods.

The results of this study can be generalized for similar occasions. Systems optimized for a specific environment can be computationally ineffective for different environments. Instead of direct usage in different environments, careful analyses and relevant performance tuning activities should be performed for adaptation.

References

[1] R.W. Burns, Communications: An International History of the Formative Years, London, Institution of Electrical Engineers, 2004.

[2] "Channel Encoding", Encyclopedia Britannica, 2010. Retrieved from Encyclopedia Britannica Online on 31 October 2010: http://www.britannica.com/EBchecked/topic/105743/channel-encoding.

[3] C.E. Shannon, "A mathematical theory of communication", Bell Systems Technical Journal, Vol. 27, pp. 379–423, 1948.

[4] R.M. Fano, "The transmission of information", Technical Report No. 65 at Research Laboratory of Electronics, Cambridge, MIT Press, 1949.

[5] D.A. Huffman, "A method for the construction of minimum-redundancy codes", Proceedings of the Institute of Radio Engineers, Vol. 40, pp. 1098–1102, 1952.

[6] Y. Çebi, G. Dalkılıç, "Turkish word n-gram analyzing algorithms for a large scale Turkish corpus – TurCo", Proceedings of the IEEE International Conference on Information Technology, Vol. 2, pp. 236–240, 2004.

[7] Ö.S. Eroğlu, "Spelling check and correction by using syllable n-gram models", MSc, Department of Computer Engineering, İstanbul Technical University, İstanbul, Turkey, 2005 (in Turkish).

[8] R. Aşlıyan, K. Günel, "Turkish automatic syllabification system and syllable statistics", Proceedings of Academic Informatics, pp. 31–38, 2008 (in Turkish).

[9] E. Çiçek, "N-gram and syllable based statistical properties of Turkish: potential applications", BSc, Department of Electronics Engineering, Ankara University, Ankara, Turkey, 2010 (in Turkish).

[10] E. Çiçek, A.E. Yılmaz, "A study on the n-gram and syllable based statistical properties of Turkish", Proceedings of the 3rd Engineering and Technology Symposium, pp. 68–77, 2010 (in Turkish).

[11] "Morse code", Wikipedia, 2010. Retrieved from Wikipedia on 31 October 2010: http://en.wikipedia.org/wiki/Morse_code.