typography,
information theory,
programming language

Marcin CHOLEWA[1]

THE READABILITY MODEL FOR NATURAL AND ARTIFICIAL LANGUAGES

This paper describes the model which allows an estimation of the readability factor of texts written in natural language or programs coded in syntax of programming languages. Only font styles are considered in this model. The destination of the model is improving readability. It can get though change font style. Several samples of text written in natural language have been used to estimation of the readability factor. Then these factors for given texts have been increased or reduced though intentional change font style. Studies have shown that deliberately changing the font style has a visible effect on improving readability or significantly lowering it.

1. INTRODUCTION

This article gives answer to question: is it possible to create and determine the parameters of the model that would describe the readability of the written text. It would take into account texts written in natural language or codes written in programming languages to program computers. This question constitutes the main thesis of this article. In the further part of the article, a model will be presented, which will have the ability to influence the readability of the text using numerical parameters. The use of such a model will improve the readability of the text. In turn, improved readability will reduce the errors caused by text reading [16]. And such errors result from problems with distinguishing neighboring characters. A better distinction between two adjacent characters guarantees their correct reading and understanding of the meaning conveyed by signs and symbols.

2. CONSTRUCT OF MODEL

The most important parameter in this model is the **readability** of the text determined by the degree of distinguishing pair of adjacent characters in the text under consideration. If the model applies to all of the text considered, all adjacent pairs of characters are considered. Afterwards, a map is created on the plane showing the readability levels for the considered text. With the help of such a map, places of greater or lesser variety of characters are visible. The generated map is a source of information about places where **readability** can be improved. It would improve by using a different font (typeface), modifying the font, using styles such as italics,

[1]University of Silesia, Institute of Computer Science, BĂŹdziÅĎska 39, 41-200 Sosnowiec, Poland
e-mail: marcin.cholewa@us.edu.pl

bolding, text size scaling, spacing between characters, or by adding or removing sheriffs. Thus, the next parameter will be the application of the new font style. It should also be said that this model does not take into account the shape of the font.

The **readability** parameter is closely related to human perception, because shows the difference in the pixel arrangement that creates the image of the character in the image being viewed [10]. The arrangement of pixels has an effect on perception [10]. The following figure explains the intuitive relationship between perception and a 2D image:

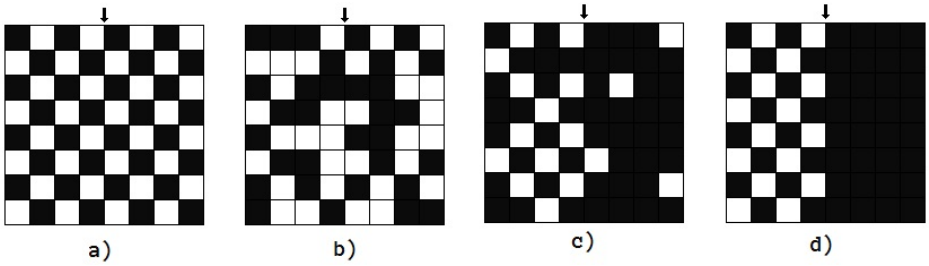

Fig. 1. The 4 chessboards with different black-and-white pools placement and the arrows of division point shown in the middle.

Figure 1 shows chessboards that show the division into areas according to the density of black or white fields. On the first two chessboards a) and b) can not see the division visually despite the division in half marked with the arrow. However, on the last two c) and d) is clearly visible.

The readability may be influenced by the distance between characters in the text. A larger space can improve readability, while a smaller one can make it much more difficult as characters will overlap. However, the analysis for the purposes of this article assumes the minimum possible spacing, such that the characters do not overlap. It is also assumed that the analysis is made of texts written in black font on a white background, because this guarantees the greatest possible contrast, which affects the readability. The font size in this study is skipped because it does not affect readability, because fonts have been used, the shape of which does not change significantly as the size increases.

Perceptually, a person can receive the first and last picture as ordered and the middle two as chaotic. In last two pictures it is easier to see the differences. In the characters it can be seen the order of pixels to a greater or lesser extent, so their ordering affects the distinction of not distinction of characters. The readability parameter is defined by the arrangement of pixel components of two adjacent characters. In the same way it will be used to study the text. The following figure shows a diagram for determining the division of positional entropy zones on which differences between pairs of characters are shown. Neighboring signs are taken into account, because the man in the process of spelling words deliberately focuses on neighboring signs because he processes the character by character. However, the human brain focuses on whole words to guess their meaning [13].

The red grid indicates the area needed for readability analysis. The size of the field in such a grid affects the accuracy of the analysis. Each field has a dimension of 8 by 8 pixels, which is enough to analyze a character with a size of 96 points. The grid height is the same for two characters in a pair, while the width is set for each character loosely, depending on the width of the character.

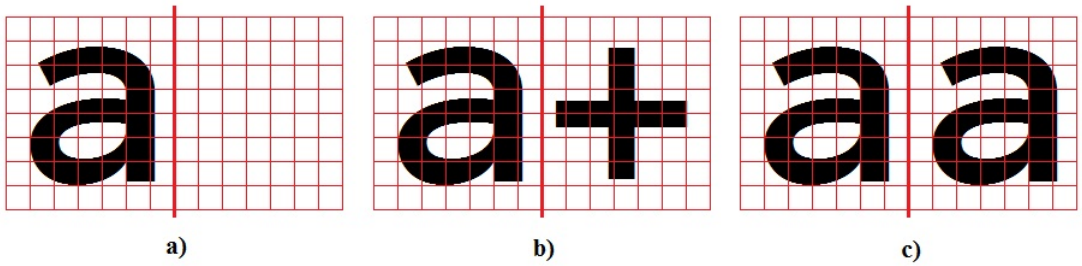

Fig. 2. In a), there are significant differences between the two characters. In c) there are no differences, as they are two identical characters.

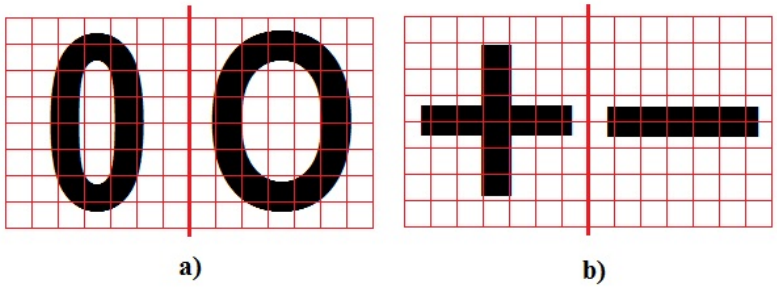

Fig. 3. Examples of character pairs in which a reading error can occur.

3. POSITIONAL ENTROPY

The *positional entropy* is a measure of readability that the author introduced. To define formal the positional entropy, there is need to be defined an auxiliary comparison function.

Definition 1. The comparison function $Eq : \Omega \rightarrow \mathbb{N}$ for two elements $x \in \Omega$ and $y \in \Omega$ is determined as follow $Eq\,(x, y) = |f\,(x) - f\,(y)|$, where $f : \Omega \rightarrow \mathbb{N}$ is simple function that returns a numeric value for its argument. What the Ω contains can be any, for example $\Omega = \{|, ||, |||\}$ and then $f\,(|) = 1$, $f\,(||) = 6$, $f\,(|||) = 2$ etc. without established order $f\,(|) < f\,(||) < f\,(|||)$.

Definition 2. The positional entropy [7] $EnpC : X \rightarrow [0, 1]$ (only for adjacent pairs) of sequence symbols $\langle a_0 a_1 a_2 a_3 ... a_{n-1}\rangle \in X$ is a measure of the degree of symbols grouping and is defined as:

$$EnpC\,(\langle a_0 a_1 a_2 ... a_{n-1}\rangle) = \frac{\sum_P \gamma\,(\{a_i, a_j\})}{Card\,(P)}$$

where:

$$P = \{\{a_0, a_1\}, \{a_1, a_2\}, \{a_2, a_3\}, ..., \{a_{n-2}, a_{n-1}\}\}$$

and γ is function determined as:

$$\gamma\,(\{a_i, a_j\}) = \begin{cases} 1, & Eq\,(a_i, a_j) \neq 0 \\ 0, & Eq\,(a_i, a_j) = 0 \end{cases}$$

for $i, j = 0, 1, ..., n-1 \wedge i = j+1$. The sequence of symbols can also be represented in the form of a rectangular matrix, so that a different definition of positional entropy can be given.

Definition 3. The positional entropy (type M) $EnpM : X \rightarrow [0,1]$ (only for orthogonally adjacent pairs) is a measure of the degree of symbols grouping in two-dimension matrix:

$$A_{m\times n} = \begin{bmatrix} a_{0,0} & a_{0,1} & \cdots & a_{0,n} \\ a_{1,0} & a_{1,1} & \cdots & a_{1,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,0} & a_{m,1} & \cdots & a_{m,n} \end{bmatrix}$$

and is defined as:

$$EnpM\left(A\right) = \frac{\sum\limits_{P} \gamma\left(\{a_{i_1,j_1}, a_{i_2,j_2}\}\right)}{Card\left(P\right)}$$

where:

$$P = \left\{ \{a_{i_1,j_1}, a_{i_2,j_2}\} : \begin{matrix} i_1 = i_2 \in [0,m] \wedge j_1, j_2 \in [0,n-1] \wedge j_1 = j_2 + 1 \\ j_1 = j_2 \in [0,n] \wedge i_1, i_2 \in [0,m-1] \wedge i_1 = i_2 + 1 \end{matrix} \right\}$$

and γ is function defined as:

$$\gamma\left(\{a_{i_1j_1}, a_{i_2j_2}\}\right) = \begin{cases} 1, & a_{i_1j_1} \neq a_{i_2j_2} \\ 0, & a_{i_1j_1} = a_{i_2j_2} \end{cases}$$

for $i, j = 0, 1, ..., n-1 \wedge i = j+1$. The function γ for $EnpM$ can be defined in a different way, such that:

$$\gamma\left(\{a_i, a_j\}\right) = \begin{cases} 1, & Eq\left(a_i, a_j\right) \in [h_1, h_2) \\ \frac{1}{2}, & Eq\left(a_i, a_j\right) \in [h_2, h_3) \\ 0, & Eq\left(a_i, a_j\right) \in [h_3, h_4] \end{cases}$$

where $i, j = 0, 1, ..., n-1 \wedge i = j+1 \wedge h_1, h_2, h_3 \in [0,1] \wedge h_1 < h_2 < h_3$. To show how this function working, the following example shows the calculation of positional entropy for the Figure 4, which can be presented in the matrix $A_{4\times4}$. Then the positional entropy will be equal to:

$$EnpM\left(A_{4\times4}\right) = \frac{\sum\limits_{j=0}^{j=4}\sum\limits_{i=0}^{i=3}\left(\{a_{j,i}, a_{j,i+1}\}\right) + \sum\limits_{j=0}^{j=4}\sum\limits_{i=0}^{i=3}\left(\{a_{i,j}, a_{i+1,j}\}\right)}{24} =$$

$$\frac{\left(1+\frac{1}{2}+\frac{1}{2}+1+\frac{1}{2}+\frac{1}{2}+1+\frac{1}{2}+\frac{1}{2}+1\right)}{24} + \frac{\left(1+\frac{1}{2}+1+\frac{1}{2}+1+\frac{1}{2}+\frac{1}{2}\right)}{24} = \frac{12}{24} + \frac{1}{2}$$

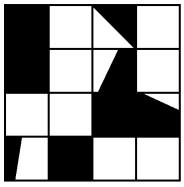

Fig. 4. The example matrix for calculating the positional entropy (M type).

For research purposes for this article, the $EnpM$ positional entropy function is used. The position entropy is similar to the Shannon entropy [15], [14], because it reaches its maximum when there are as many as possible adjacent pairs of different characters in sequence. The positional entropies ($EnpM$) for the halves of cubes from Figure 1 are as follows: $EnpM\left(a_1\right) = 1$,

$EnpM\,(a_2) = 1$, $EnpM\,(b_1) = 0.635$, $EnpM\,(b_2) = 0.692$, $EnpM\,(c_1) = 1$, $EnpM\,(c_2) = 1$, $EnpM\,(d_1) = 0.6$, $EnpM\,(d_2) = 0.5$.

The following example shows how the positional entropy value changes when the 0 character of the font is modified to make it more readable. For the Figure 5. a), the difference of the positional entropy of the characters 0 and o is $EnpM_1 = |0.148 - 0.153| = 0.005$. For the Figure 5. b) is $EnpM_2 = |0.148 - 0.131| = 0.017$. It can be clearly seen that $EnpM_1 < EnpM_2$, which says unequivocally that the change of the 0 character affects better readability.

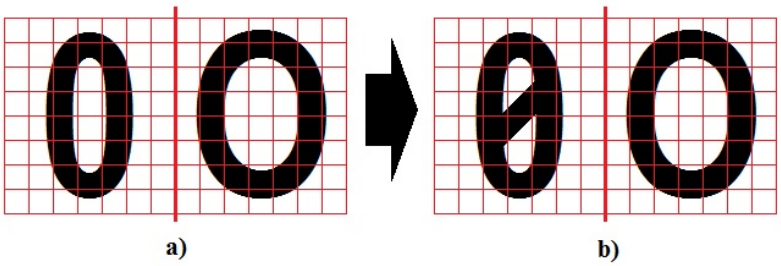

Fig. 5. The effect of modification of the 0 character (digit zero) to improve its distinction from the o character (letter o).

Definition 4. Each character sequence $\langle a_0 a_1 a_2 a_3 ... a_{n-1} \rangle$ with n-length that participates in the readability analysis will be presented in the form of readability map. The map presents the degree of readability between two adjacent characters for all character pairs. Each point $p_0, p_1, ..., p_{\lceil \frac{n-1}{2} \rceil}$ on this map presents the degree of readability for pairs of adjacent characters.

The map must have the shape of a rectangle or a square depending on the number of p points. If the square root of the number of all p points is an integer, then the map has the shape of a square. If the number of p points is a prime number, add an additional p point with a value of 0 to create a full rectangle.

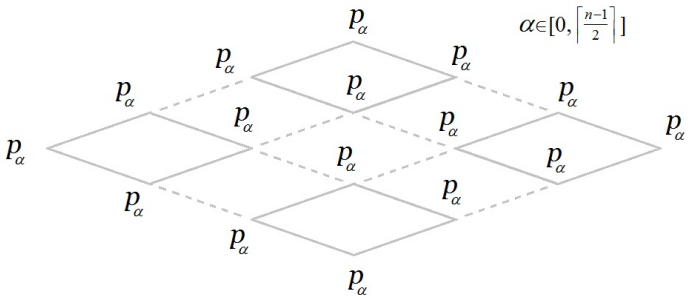

Fig. 6. The readability map of the character string with n characters.

Definition 5. For character sequence $\langle a_0 a_1 a_2 a_3 ... a_{n-1} \rangle$ with n characters, the total readability parameter is defined as:

$$c = \frac{\sum\limits_{k=1}^{m} \left(d\left(p_1, p_2\right) + s\left(p_1\right) + s\left(p_2\right) \right)}{\sum\limits_{k=1}^{M} d\left(p_1, p_2\right) + 2M}$$

where $d\left(p_1, p_2\right)$ is a distance between p_1, p_2 using taxi metric, $s\left(p_1\right)$, $s\left(p_2\right)$ are degree of readability for given pairs. It is assumed that $m < M$.

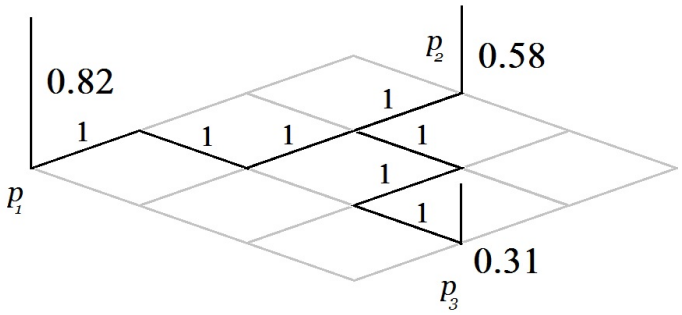

Fig. 7. The visualization of the formula from Definition 4 based on the readability map for p_1,p_2,p_3 points. In additional, it is assumed that $m = 3$, $M = 16$, $d(p_1, p_2) = 4$, $d(p_2, p_3) = 4$, $s(p_1) = 0.82$, $s(p_2) = 0.58$ and $s(p_3) = 0.31$.

4. ANALYSIS AND CONCLUSIONS

In the first readability analysis, a 236-character natural language text was used. The text has been written in different fonts, and then for each font the style has been changed (italic, bold, italic-bold)[8] to see how the readability factor changes. These styles have been used, because in this way it can modify font styles without interfering with font structure. After the analysis of each font, a readability map for the given font was generated to calculate the total readability parameter.

Table 1. The results of the analysis of the readability of text in natural language written in different font styles.

Typeface (before)	**Total readability (before)**	**Style (after)**	**Total readability (after)**	**Improving readability [%]**
Times New Roman	0.8086	Times New Roman, italic	0.8053	-1%
Times New Roman	0.8086	Times New Roman, bold	0.8670	6%
Times New Roman	**0.8086**	**Times New Roman, italic, bold**	**0.8864**	**8%**
Times New Roman Cond.	0.8329	Times New Roman Cond., italic	0.8610	3%
Times New Roman Cond.	0.8329	Times New Roman Cond., bold	0.8529	2%
Times New Roman Cond.	0.8329	Times New Roman Cond., italic, bold	0.8448	1%
Arial	0.8735	Arial, italic	0.9108	4%
Arial	0.8735	Arial, bold	0.9079	3%
Arial	0.8735	Arial, italic, bold	0.8833	1%
Courier New	0.8761	Courier New, italic	0.8753	-1%
Courier New	0.8761	Courier New, bold	0.8560	-2%
Courier New	0.8761	Courier New, italic, bold	0.7283	-15%
Tablica	0.9180	Tablica, italic	0.8012	-11%
Tablica	0.9180	Tablica, bold	0.8426	-8%
Tablica	0.9180	Tablica, italic, bold	0.7667	-15%
Orbit	0.9200	Orbit, italic	0.7282	-19%
Orbit	0.9200	Orbit, bold	0.8792	-4%
Orbit	0.9200	Orbit, italic, bold	0.6734	-25%

The most important column in Table 1 is *Improving readability* and it tells if the change in font style improved the readability factor (when positive value). The greater the readability factor, the more readable the text for a given font. This means that for a given font, the text contains fewer pairs of characters that can cause errors in reading by human. The Times New Roman [9] and Times New Roman Condensed fonts increase the readability after style change. For the Arial font is similar. The conclusion is clear because these styles can be used for these fonts to highlight important information and protect from reading errors. There is a significant deterioration in readability for the Tablica and Orbit fonts. The Orbit font is futuristic [5] and has a high readability factor, however, without modifying the style. It is also worth emphasizing

that the Tablica font also has a high readability factor and is used in direction tables in the Polish road system [12], [11]. That factor can be would modify to increase the readability factor. The greatest improvement was found for the italic-bold style of the Times New Roman font, because the sheriffs in this font affect the readability caused by distinguishing characters. On the other hand, readability for the Times New Roman font without modification is the smallest in the surveyed group.

In the second analysis, the source code of the 288-character algorithm was used. This algorithm was presented using several known fonts used by programmers[6], [1], [2].

Table 2. The results of the readability analysis of the source code of the algorithm written in different font styles.

Typeface	Total readability
Times New Roman	0.8638
Courier New	0.9264
Consolas	0.9218
Complex	0.9076
Terminal	0.8988
Anonymous	0.9353
IBM 3270	**0.9760**
Hack	0.9251
Liberation	0.8804
Envy R	0.9713

It turned out that the most readable font for the source code was the screen font of the IBM 3270 computer terminal produced in the 1970s [3]. Geometric shapes uniquely mark out individual letters. Additional characters such as commas, periods, strips, percentages and symbols of mathematical operators are well visible in the source code. In each of the tested fonts, a b o pair was found which significantly reduces readability. The shape of the letter b or o should be redesigned. For comparison, the Times New Roman font is also chosen, whose characters do not have equal width. It turned out that for this font the readability is the smallest, so it is not suitable for editing the source code. The Liberation font, which developers use, in terms of total readability is a bit better than Times New Roman and it should not be used.

The purpose of the third analysis was to determine the readability factor for a string containing only the digits. The string of digits has been generated so that all possible pairs for numbers appear.

Table 3. The results of the readability analysis of strings composed only of digits written in different font styles.

Typeface	Total readability
Times New Roman	0.9282
Courier New	0.8781
Consolas	0.9018
Complex	0.8297
Terminal	0.8712
Anonymous	0.9224
IBM 3270	0.9040
Hack	**0.9422**
Liberation	0.8553
Envy R	0.9178
Tablica	0.7949

An interesting fact is that for the Tablica font, the smallest readability value for the string of digits was obtained, and should be the reverse situation. The Hack font[2] has the highest

readability, which guarantees fewer mistakes in reading numbers in the algorithm code. For most of the fonts included in the analysis, the problem in reading is for pairs of digits 0 8, 5 2, 3 5, 2 3, 9 6. In these pairs, it's easiest to make a mistake between the digits. The characters in these fonts should be modified to get better readability. In the last analysis, the IBM 3270 font for the numbers themselves does not perform as well as the readability of the source code.

5. FURTHER WORK

The conclusions after the analysis suggest the creation of a new font, more resistant to the generation of reading errors caused by similarity to the adjacent character in the string. Although a font has been created that facilitates reading the text by people with dyslexia [4]. In the future, such a model could be an element of a system that generates the optimal syntax of program languages due to smaller human errors when writing a program. It should also be confirmed that the presented results, by examining a group of people, are statistically more errors in reading the correct characters in places indicated by the model. The surveyed people should be divided into healthy people and struggling with dyslexia [16].

BIBLIOGRAPHY

[1] ANONYMOUS PRO FONT. https://www.marksimonson.com/fonts/view/anonymous-pro [Online; Accessed: 6.11.2018].
[2] HACK FONT. https://sourcefoundry.org/hack [Online; Accessed: 8.11.2018].
[3] HISTORY OF IBM 3270. http://www.ibm.com/support/knowledgecenter/en/SSGMGV_3.1.0/com.ibm.cics.ts31.doc/dfhp3/dfhp3bh.htm [Online; Accessed: 12.11.2018].
[4] OPENDYSLEXIC MONO FONT. https://www.opendyslexic.org [Online; Accessed: 6.11.2018].
[5] ORBIT-B FONT. https://www.myfonts.com/fonts/bitstream/orbit-b [Online; Accessed: 6.11.2018].
[6] THE BEST PROGRAMMER FONTS. https://speckyboy.com/best-free-fonts-coding[Online; Accessed: 10.11.2018].
[7] CHOLEWA M., PALYS M. Estimation of information entropy based on its visualization. Journal of Medical Informatics & Technologies, 2017, Vol. 26. pp. 18–25.
[8] CLAYTON E. The golden thread: The story of writing. Atlantic Books, 2013. pp. 104–6.
[9] DREYFUS J. The evolution of times new roman. The Penrose Annual, 1973, Vol. 66. pp. 165–174.
[10] GILCHRIST A. Seeing black and white. Oxford Psychology Series, 2006.
[11] MINISTERSTWO KOMUNIKACJI. Instrukcja o znakach i sygnałach na drogach. Wydawnictwo Komunikacji i Łączności, 1975.
[12] MISIAK M. Pismo drogowe. FUTU Paper, 2013, Vol. 10. p. 25.
[13] PATERSON K. B., READ J., MCGOWAN V. A., JORDAN T. R. Children and adults both see 'pirates' in 'parties': Letter-position effects for developing readers and skilled adult readers. Developmental Science, 2014, Vol. 18 no. 2.
[14] PIERCE J. R. An introduction to information theory: Symbols, signals, and noise. Dover Publications, 1980.
[15] SHANNON C. E. A mathematical theory of communication. Bell System Tech. J., 1948, Vol. 27 no. 3. pp. 379–423.
[16] ZIEGLER J. C., PECH-GEORGEL, C. DUFAU S., GRAINGER J. Rapid processing of letters, digits and symbols: What purely visual-attentional deficit in developmental dyslexia? Developmental Science, 2010, Vol. 13 no. 4. pp. F8–F14.