A method for language attribution based on assessment of text irregularity

Cite as: AIP Conference Proceedings **1982**, 020006 (2018); https://doi.org/10.1063/1.5045412
Published Online: 30 July 2018

Dmitry A. Tarasov

A method for language attribution based on assessment of text irregularity

Dmitry A. Tarasov

Ural Federal University, Ekaterinburg, Russia

datarasov@yandex.ru

Abstract. In the work, it is proposed to use a fractal driven index irregularity as a base for language attribution of text. A method for such an attribution of electronic texts is offered. The method may be applied in Big Data analysis, electronic library operation, natural language processing and so on. We examined texts in nine languages, set up with two major fonts. The approach shows the ability to distinguish different languages using the offered irregularity based index only, without reading texts and expert language assessment. The method may be further extended for texts in bitmap.

INTRODUCTION

Current intensive grow of information requires special tools for data assessment. Analysis of text data is one of the major areas of information processing in Big Data applications. Existing methods for textual information analysis take into account only its semantic part and do not operate with its spatial form [1]. It is known that the level of understating of the textual materials depends on a font's spatial form [2], [3]. However, for a long period it had not been proposed a method for a numerical assessment of spatial features of texts. Using irregularity of fonts as quantitative assessment of the font's drawing allows expand a set of measurable parameters at textual data analysis [4]. The proposed irregularity is a fractal driven and scale invariant index accounting spatial features of a particular font. The method of irregularity (C) calculation is simple and do not require high computing power [4]. The value of the indicator is calculated by the formula (1).

$$C = \frac{P^2}{4\pi S}, \tag{1}$$

where P is a total perimeter of curves, and S is a total area of characters from the set of letters that forms the font. As a set of letters, we used all uppercase and all lowercase letters from the Russian (or any other alphabetically based) language, thus the irregularity depends not only on the font's shape but also on number of letters in a language.

For the calculations, we utilized vector forms of fonts, which were operated in the CorelDraw vector software package. Further, we confirmed additivity of the irregularity [5] and proposed a method for calculation of irregularity for raster fonts by their bitmaps [6]. Thus, the method might be considered sufficiently developed.

The obvious modification of the approach might imply specific units of the irregularity. First, we might divide the irregularity by the number of characters. Moreover, to determine the quantitative characteristics of the text, it does not necessarily use only the full set of letters in the language. We can apply the calculation of irregularity for any set of characters that is large enough to represent the language. The volume of such a sufficient set is still a question. Finally, we can simplify the irregularity itself by removing the constant 4π from consideration and by modifying the formula (1) and the calculating method without significant impact on the result. The purpose of such a modification might be application of the approach for text attribution in order to distinguish different languages. As each language has its unique hidden, intrinsic feature that can be defined by the text structure and letters' frequency

Mathematical Methods and Computational Techniques in Science and Engineering II
AIP Conf. Proc. 1982, 020006-1–020006-6; https://doi.org/10.1063/1.5045412
Published by AIP Publishing. 978-0-7354-1698-7/$30.00

appearance analysis [7], we can expect that any particular language might has its own irregularity-based index, which defines the language in a unique way.

The aim of the work is to offer the method for language attribution based on assessment of text irregularity.

APPROACH

Since the language as a whole is characterized by a particular index (based on χ^2 statistics or other nature), each set of letters forming a coherent text, large enough to represent the language might be considered a language "unit". The task is to distinguish a numerical value of this "unit".

We propose to use an irregularity-based index, which we call I factor (2).

$$I = \frac{P^2}{Sn}, \tag{2}$$

where n is a number of characters in the text excerpt being assessed. Thus, we can say that the index is the average character irregularity of the font of the chosen language.

To calculate such a factor, we use the same method as used for irregularity assessment applying CurveInfo macro in CorelDraw application [4]. The only limit is that this method is applicable for vector representations of texts only. However, further it is easy to expand the bitmaps assessment technique [6] for such a calculation.

For the experiment, we selected two major fonts (Arial and Times New Roman) and nine European alphabetically based languages (Russian-Rus, English-Eng, French-Fra, Italian-Ita, Spanish-Spa, German-Ger, Turkish-Tur, Greek-Gre, and Czech-Cze). We also assessed the *I* factor for highly irregular alphabets of Georgian-Geo, Armenian-Arm and Arabic-Ara languages.

As a text sample, we used the famous novel "War and Peace" by Leo Tolstoy as it was easy to find this piece of art in different languages.

The first stage of the experiment consisted of *I* factor calculations for each selected language (including three additional languages for comparison) and for two selected fonts. For this stage, we used text excerpts of 1000-2000 characters.

The second stage was ranking languages by their *I* factors and building dependency graphs.

The third stage of the work was to identify the behavior of the *I* factor when changing the number of characters in the evaluation sample and plot dependency graphs. For this stage, we used Arial font only as behavior of *I* factors for Times New Roman is the same.

RESULTS AND DISCUSSIONS

The results of *I* factor calculations for nine basic and three additional languages and for two fonts are shown in Table 1 and in Figure 1. The languages have been already ranked by increase of I factors.

As it may be seen from the Table 1, the additional languages (Georgian, Armenian and Arabic) have significantly higher I factors than other ones, so further we eliminated them from consideration.

In Figure 1, Arial and Times New Roman outlines for Arabic language (Ara) were the same.

The behavior of the *I* factors for the Arial font for nine languages when changing the number of characters in the evaluation text samples are shown in Table 2 and Figure 2.

As it can be seen from the Figure 2, different languages demonstrate a substantially different behavior of the *I* factor. The relative stability of the dependency behavior begins to be observed after approximately 3,000 characters in the text sample. Moreover, we can predict that such a behavior may significantly vary when using different fonts. Thus, the predictor of a language can be not only the factor itself, but also the dynamics of its change calculated for different fonts.

We see the opportunity to distinguish between languages by the *I* factor, and by its dynamics. The exception may be close languages, such as Italian and Spanish, where the difference can be traced only by the dynamics of the *I* factor, as it can be seen in Table 2 and in Figure 2.

TABLE 1. . *I* factors for different languages.

Language	Number of characters in samples	*I* factor for Arial	*I* factor for Times New Roman
Greek - Gre	1145	41.913	53.226
Italian - Ita	2168	44.621	54.872
Spanish - Spa	1133	44.913	55.385
French - Fra	1721	46.920	58.227
English - Eng	979	47.762	60.709
Czech - Cze	947	48.081	60.771
Turkish - Tur	1077	48.573	61.760
Russian - Rus	1009	50.115	66.612
German - Ger	1313	52.108	66.341
Georgian - Geo	924	61.061	75.382
Arabic - Ara	808	65.839	65.834
Armenian - Arm	1023	88.075	97.970

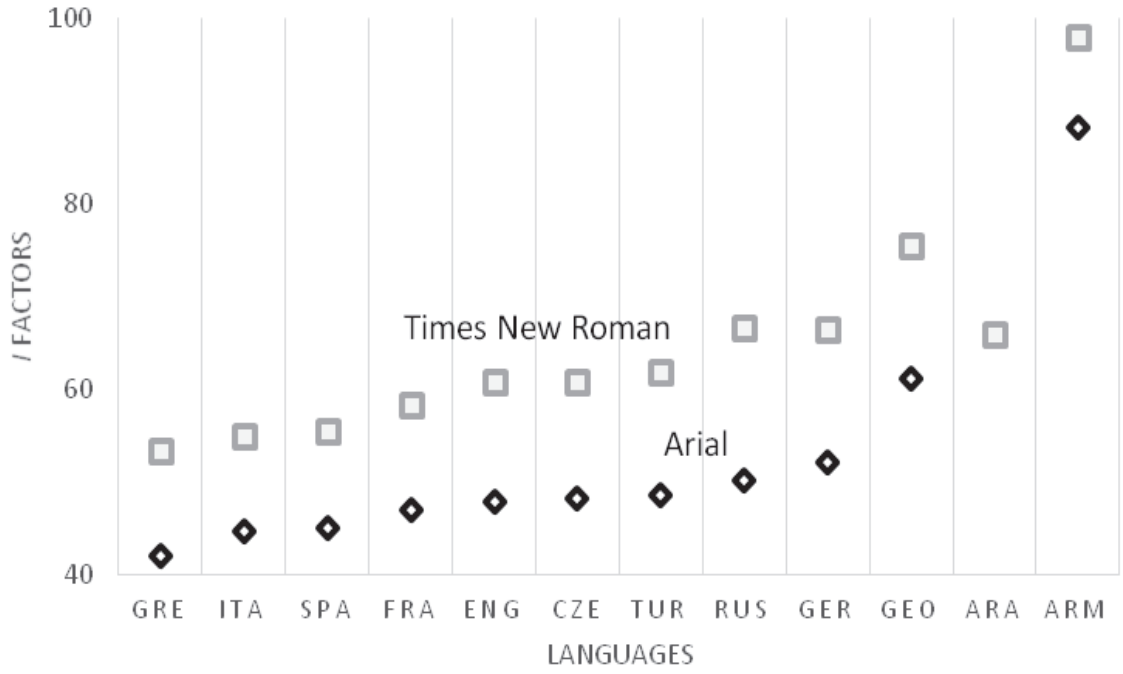

FIGURE 1. I factors for twelve languages and two fonts

TABLE 2. . *I* factors dynamics for different languages.

	I factors for different ranges of number of characters in a text sample						
Language	20-50	80-160	180-300	400-600	700-900	901-2200	2201-4000
Greek - Gre	42.114	42.223	42.239	41.869	41.482	41.913	42.065
Italian - Ita	46.442	44.799	44.740	44.915	45.613	44.621	45.414
Spanish - Spa	44.100	45.088	44.809	44.965	44.762	44.913	45.124
French - Fra	39.567	43.636	44.955	45.509	46.185	46.920	47.635
English - Eng	45.972	50.380	49.644	47.808	47.309	47.762	48.306
Czech - Cze	44.467	47.339	47.510	48.867	48.074	48.081	49.200
Turkish - Tur	49.247	49.040	48.723	48.468	47.809	48.573	50.298
Russian - Rus	60.186	51.211	48.289	50.155	49.734	50.115	48.543
German - Ger	56.613	54.555	53.496	52.503	51.824	52.108	51.751

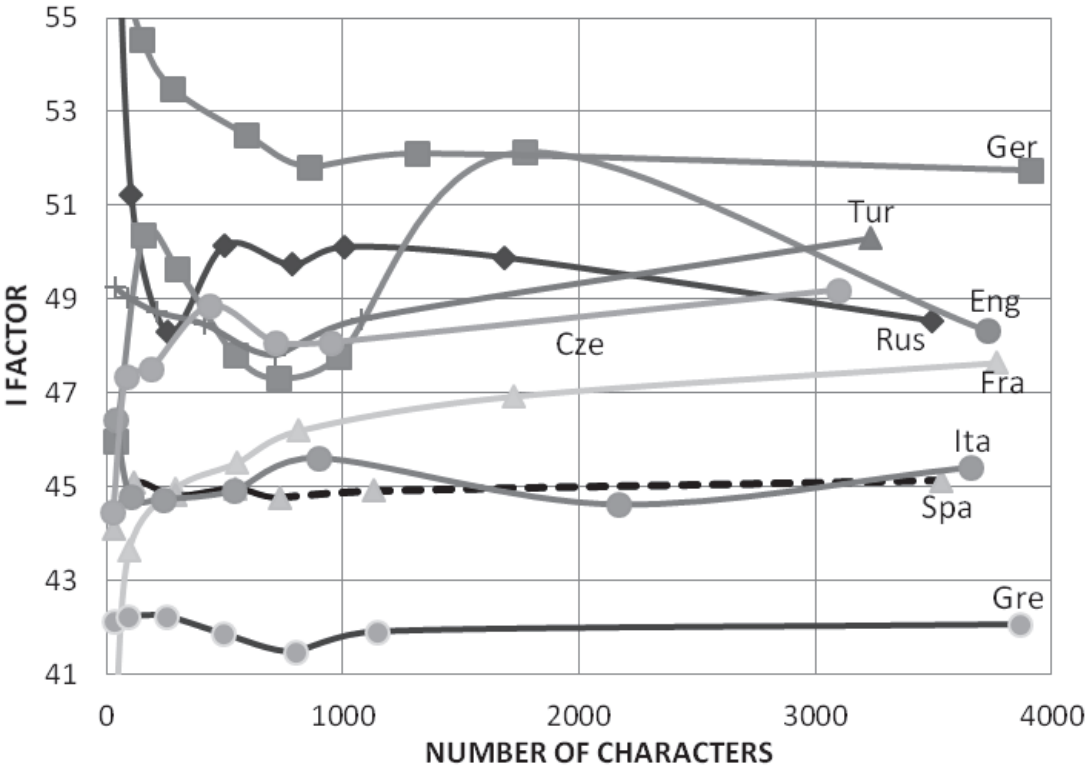

FIGURE 2. I factors behavior for different languages depending on the number of characters, Arial font

The extension of amount of characters in the text sample for I factor calculation may lead to stabilization of the factor. This may be utilized in Big Data operation as an additional document attribute that expands the ability to sort documents and improve the work of search engines.

Investigations of the dynamics of the factor will allow eliciting the structural features of different languages, which can be used in structural linguistics and analysis of unknown languages.

In the more complicated case of raster text samples (scanned documents, manuscripts, books etc.) we can apply the approach [6] that requires modification just as it is described above. The only problem might be scan quality as correct irregularity calculation strongly depends on the resolution of a rasterized sample.

CONCLUSION

Big Data operation, particularly in current search engines requires fast approach for text attribution in the light of the extraordinary growth in the number of documents available for analysis.

Thousands of studies in legibility have led to contradictory results. Moreover, until now, there is no consensus among scientists about what factors and how affect reading and spatial text perception. Researchers do not know how account the spatial form of the text as in legibility studies, so in numerical text analysis.

Currently, it was possible to measure almost all spatial features of text except font. This was probably due to the lack of an objective index, which could describe a typeface numerically. If we numerically describe font we can describe text as a whole.

We assumed that the similarity of some graphic elements of letters in font and the letters themselves, as well as the font as a whole, suggests the possibility of using the ideas of fractal geometry to make such an assessment. Fractal dimension that can be understood as the degree of filling the space by an irregularly distributed substance. The fractal Minkowski dimension d combines the number of objects n and their geometric size a. Mandelbrot further showed that for fractal sets the expression related to the length of the perimeter P and the area S of the object. In other words, in either family of flat figures (e.g. characters of a font), that are geometrically similar but having different linear dimensions, the ratio of the length of the shapes border to the square root of its area is a number that is completely determined by the general form for the family. Thus, we defined the compactness and irregularity for a set of characters. The set of characters in our works is represented by the whole set of font's letters together with its division into internal and external volumes because the account of these volumes in the formula is made differently.

Application the irregularity as a quantitative assessment of the font's drawing allows to expand a set of measurable parameters at textual data analysis. The irregularity is approved to be scale invariant. We have also confirmed its additivity and the negative correlation between irregularity and reading speed.

The proposed method of irregularity assessment is simple and do not require high computing power. For the calculations, we utilized vector forms of fonts, which were operated in the CorelDraw software package with help of CurveInfo macro. Further, we proposed a method for calculation of irregularity for raster fonts by their bitmaps.

In this work, we offer a simple method for language attribution of electronic texts by a specially developed I factor. The factor, as well as its forerunner irregularity, is based on the ideas of fractal geometry and may be easily implemented in any text analysis system.

As each language has its unique hidden, intrinsic feature that can be defined by the text structure and letters' frequency appearance analysis, we can expect that any particular language might has its own irregularity-based index, which defines the language in a unique way. Ihe index is the average character irregularity of a particular font of the chosen language. To calculate such a factor, we use the same method as used for irregularity assessment.

The possibility of extension of the method onto the field of rasterized text (scanned texts, documents, manuscripts, books) stored in electronic libraries may affect to the improvement of library documents management.

The proposed approach might be applied together with other methods of language and documents processing and is able to lead to their further development.

REFERENCES

1. G. Amir, H. Murtaza, "Big data concepts, methods and analytics". International Journal of Information Management, 2015, 35, p.140.
2. K. Larson, "Measuring the Aestetics of Reading". People and computers XX. Engage: proceedings of HCI 2006, the 20nd British HCI Group annual conference. UK, 2007, pp. 41–56.
3. D. Tarasov, Vision and reading (Зрение и чтение). Ekaterinburg: UrFU, 2015, ch. 3. (in Russian)

4. D. Tarasov, A. Sergeev, “Irregularity as a quantitative assessment of font’s drawing and its effect on the reading speed”. CEUR Workshop Proceedings. Supplementary Proceedings of the 4th International Conference on Analysis of Images, Social Networks and Texts (AIST'2015). 2015. Vol.1452. 177-182.
5. D. A. Тарасов, A. S. Sydikhov, A. P. Sergeev, A. G. Tyagunov “Additivity of irregularity of outline fonts (Аддитивность изрезанности контурных шрифтов)”, Proceedinds of International conference «Information: transfer, operation, Perception», Ekaterinburg, UrFU. 2016, pp. 4-19. (in Russian)
6. D. A. Тарасов, A. P. Sergeev, A. G. Tyagunov, “Assessment of ireegularity of a raster font by its bitmap image (Оценка изрезанности растрового шрифта по его битовому изображению)”, Proceedings of the higher educational institutions. Problems printing and publishing, 2015, № 3, pp.60-67. (in Russian)
7. V. V. Filimonov, A. M. Amieva, A. P. Sergeev “Clustering of Russian-language texts using χ^2 statistics (Кластеризация русскоязычных текстов с применением статистики χ^2)”, Proceedinds of International conference «Information: transfer, operation, Perception», Ekaterinburg, UrFU. 2016, pp. 164-174. (in Russian)