Language Attribution of an Unmarked Text Corpus

DMITRY TARASOV
Department of IT and Automation
Ural Federal University
Mira 32 – R041, Ekaterinburg 620002
RUSSIA
datarasov@yandex.ru http://www.urfu.ru

Abstract: Unmarked text corps will increasingly appear with the growth of information on the web. Automated analysis of Big Data in search engines, scientific and commercial applications requires detailed information about the object under study. In the case of text bodies, information on the language of the documents is extremely important. Working with the scanned texts the situation is even more complicated. In this paper, the idea of using the fractal-inspired *irregularity* to attribute the language of the text is being further developed. A methodology for the attribution is proposed and an experiment based on 10 European languages is conducted. The proposed approach has shown its effectiveness and promise. A selection of approximately 4000 characters (1 page of text) allows you to uniquely attribute the language of the text.

Key-Words: - Big Data, Fractal, Irregularity, Language

Received: May 20, 2020. Revised: November 27, 2020. Accepted: December 11, 2020. Published: December 29, 2020.

1 Introduction

The analysis of large text data arrays is one of the main areas of information processing in Big Data applications oriented to web search engines, as well as, to the applied linguistics and related fields of scientific search. Existing methods of textual information analysis operating with a various corpus of texts take into account only its semantic part and do not operate on its spatial form [1]. Nevertheless, it was found that the level of understanding of the text depends, in particular, on the spatial characteristics (shape, size, etc.) of the font used [2, 3]. However, this fact is usually not taken into account when analyzing texts, and for a long period, no method was proposed for numerically assessing the spatial features of texts.

Legibility and understanding of the text have been studied for over a hundred years. One of the main objects in these studies is the font. Many researchers investigated various features of the font, such as legibility, readability, the influence of serifs, the influence of the pattern and spatial characteristics of the font on understanding and remembering of the text content, and some other factors. An almost equal number of studies showed the advantages and disadvantages of serifs, as well as the preference for other spatial features of the text. The preferences of certain font features and font sizes have also diverged greatly. It can be assumed that legibility is more sensitive to some combinations of spatial features of the text. No special font is highlighted as the most convenient for use in publishing. The point to pay attention to is the familiarity of respondents with a particular typeface and subjects' preferences.

Thus, thousands of studies did not reveal the truth and led only to conflicting results [4]. Moreover, there is still no consensus among scientists about what factors and how to affect reading, in particular, the perception of a spatial text. This is probably due to the lack of an objective criterion that could describe the font quantitatively. Scientists cannot quantify font patterns and therefore cannot objectively compare different fonts. Assessing the visual properties of a font presents certain difficulties associated with a misunderstanding of what constitutes a set of visual font functions and what assessment criteria should be used. The similarity of some graphic elements of letters in the font and the letters themselves, as well as the font as a whole, suggests the possibility of using the ideas of fractal geometry for such an assessment.

This study aims to expand the method for language attribution based on the assessment of the fractal-driven text *irregularity* and define the criteria for the automated text language attribution based on the irregularity.

2 Approach and Experimantal

We hypothesized that the prospected method for numerical evaluation of fonts is to consider the fractal dimension, which can be understood as the degree of filling the space with an unevenly distributed substance. The Minkowski fractal dimension d is expressed by formula (1), which combines the number of objects n and their geometric size a.

Mandelbrot further [5] showed that for fractal sets the expression related to the length of the perimeter P and the area S of the object is performed by equation (2).

$$d = \lim_{a \to 0} \frac{\log n(a)}{\log \frac{1}{a}} \tag{1}$$

$$P^{\frac{1}{d}} \Big/ S^{\frac{1}{2}} = const, \tag{2}$$

which implies that $S \sim P^{\frac{2}{d}}$

Thus, in either family of flat figures (*e.g.*, characters set of a font) that are geometrically similar but having different linear dimensions the ratio of the length of the shape's border to the square root of its area is a number that is completely determined by the general form for this family. Therefore, one may define the *compactness* for a set of characters (*e.g.*, a complete set of characters in a given language for a particular font, see Fig. 1, above) as the relation between the contour length (P) of the set and its area (S). 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. The sample of such a division for the Russian language (33 uppercase and 33 lowercase letters, font Arial) is shown in Fig. 1 (below).

Fig. 1. Set of characters (above) and its division (below) for the *irregularity* calculation; Russian alphabet, font Arial

To obtain information about the perimeter of a particular character (letter), the perimeter of the external contour of a letter (P_{out}) must be added to (if available) the internal perimeter of the letter space (P_{in}). The perimeter of the whole set (P) equals the sum of the perimeters of all letters (3). To obtain the area of a letter it must subtract the internal area of the letter space (S_{in}) from the general area bounded by the outer contour of the letter (S_{out}). The area of the whole set (S) is equal to the sum of the areas of all the letters (4).

$$P = \sum(P_{out} + P_{in}) \tag{3}$$

$$S = \sum(S_{out} - S_{in}) \tag{4}$$

The concept of the *irregularity* (C) is introduced by the same way (5).

$$C = P^2 \Big/ 4\pi S, \tag{5}$$

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* introduced as described depends not only on the font's shape but also on the number of letters in a language.

Application of the *irregularity* as a quantitative assessment of the font's drawing [6] allows expanding a set of measurable parameters of text that might be utilized in textual data analysis. The proposed *irregularity* is a fractal-driven index account for the spatial features of a particular font and, probably, a text as a whole. The index is approved to be scale-invariant. Further, the additivity of the *irregularity* was confirmed [7]. The correlation between the *irregularity* and reading speed was observed. The method of the *irregularity* calculation is simple and does not require high computing power. For the calculations (3)–(5), vector forms of fonts are utilized. They are operated in the CorelDraw vector software package with help of CurveInfo macro after the division of a font sample into primitives as described in Fig. 1.

A method for calculation of the *irregularity* for raster fonts by their bitmaps was also proposed [8]. This might be of particular interest to those who need to assess paper (and scanned) texts. Since the calculation of perimeters and areas of the rasterized objects is methodologically complicated, the application of the approach in the case of scanned texts has to be modified.

First, we suggest using 1-bit bitmap forms of samples (Fig. 1). Second, we need to distinguish a 1-pixel border of each character and the whole sample (Fig. 2). This procedure is easily processed

by applying raster image operating software (e.g. Corel PhotoPaint, Adobe Photoshop, *etc.*).

АБВГДЕЁЖЗИЙКЛМНОПРСТУФХЦЧШЩЪЫЬЭЮЯ
абвгдеёжзийклмнопрстуфхцчшщъыьэюя

Fig. 2. Set of characters for the bitmaps *irregularity* calculation; Russian alphabet, font Arial, 1-pixel border, 150 ppi

The 1-pixel border acts as a perimeter in the case of bitmaps, as well as, an area is calculated also in pixels. Since in the case of bitmaps the area and 1-pixel border depend on bitmap's resolution, this is a prompt to use a fractal approach, too, therewith a calculated 1-pixel border is none other than the fractal's border area. The geometric size a in (1) is also directly proportional to the resolution of bitmaps.

In order to compute the *irregularity* of a bitmap font image, it is necessary to calculate the number of pixels in the sample (e.g., Fig. 1, above) and sample's border (e.g., Fig. 2) that might be done in practically every kind of the image processing software. Obviously, the higher the resolution of the bitmap, the more accurate assessment of the *irregularity* is achieved.

The plots of dependencies of areas and 1-pixel borders from the parameter $1/a$ for different resolutions revealed their characteristic behavior. The areas depend on the parameter $1/a$ quadratically (see Fig. 3, above), and perimeters depend on it linearly (see Fig. 3, below), which also confirms the fractal nature of the *irregularity*.

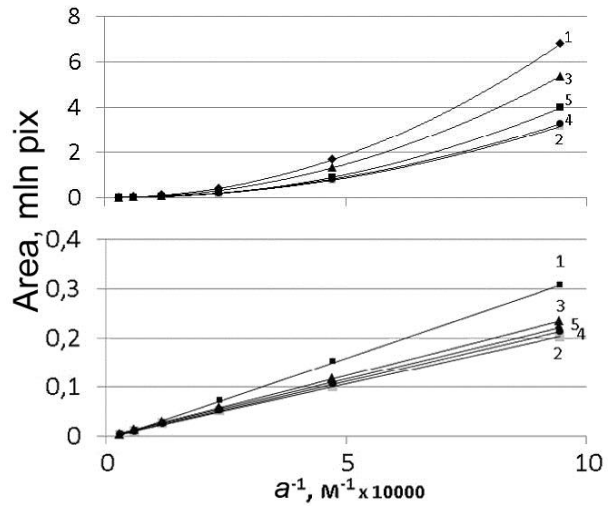

Fig. 3. Dependencies of the number of pixels in areas (above) and 1-pixel borders (below) from the 1/a fractal parameter for different fonts: 1) Arial, 2) Futuris, 3) Times New Roman, 4) Classic Russian, 5) Art Script

Thus, the method of the text *irregularity* assessment might be applied for both vector and bitmap representations of fonts and texts, although the algorithms for the *irregularity* calculating are significantly different. Moreover, the automation of the bitmap *irregularity* assessment seems pretty complicated, whilst the vector form of the *irregularity* is easier to operate and might be incorporated into specialized software (for instance, by processing PDF documents).

The next stage of modification of the approach is implying specific units of the *irregularity*. In order to determine the quantitative characteristics of the text, it does not necessary to use the whole set of letters in the language. We can apply the calculation of the *irregularity* for any set of characters that is large enough to represent the language.

Since the language as a whole might be characterized by a particular statistical index [9], each set of letters forming a coherent text, large enough to represent the language might be considered a language "unit". It is known that each language has its unique hidden, intrinsic feature that can be defined by the text structure and letters' frequency appearance analysis. Thus, we may expect that the *irregularity* somehow might reveal such a structure and define the language of a text in a unique way.

In the previous work [10], it is proposed using the *irregularity*-based index, which is called *I* factor (6).

$$I = {P^2}/{Sn} \tag{6}$$

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 a particular font of the chosen language.

To calculate such a factor (in a vector form only, yet), the same method as used for the *irregularity* assessment applying CurveInfo macro in CorelDraw application [6] is used. The bitmaps assessment technique is not yet developed enough.

Thus, the *I* factor has been calculated for two widely used document fonts (Arial and Times New Roman) and nine European alphabetically based languages (Russian-Rus, English-Eng, French-Fra, Italian-Ita, Spain-Spa, German-Ger, Turkey-Tur, Greek-Gre, and Czech-Cze) (see Table 1). Table 1 also contains the number of characters in the corpus (sample). Expanding the font list is a future challenge. At the moment we are interested in those fonts that are used in the vast majority of documents.

Further, during the experiments, the volume of sufficient character set defining the language is considered not less than 2000 letters (see Fig. 4). It was obvious that the dependences of the *I* factor on the number of characters in the corpus become approximately constant when a value of about 2000 is reached.

TABLE I. THE RESULTS OF THE *I* FACTOR CALCULATION FOR THE COMMON ARIAL AND TIMES NEW ROMAN FONTS AND NINE EUROPEAN LANGUAGES

Language	*Characters in the corpus*	*I facror (Arial)*	*I facror (Times New Roman)*
Greece	1145	41,91	53,23
Italian	2168	44,62	54,87
Spanish	1133	44,91	55,39
Franch	1721	46,92	58,23
English	979	47,76	60,71
Czech	947	48,08	60,77
Turkish	1077	48,57	61,76
Russian	1009	50,12	66,61
German	1313	52,11	66,34

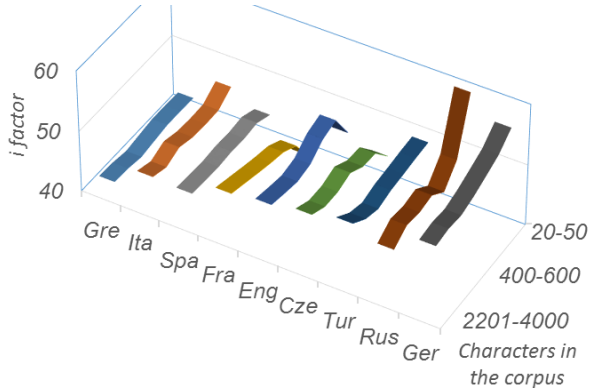

Fig. 4. I factors change depending the volume of the corpus

However, in the previous study, quantitative relationships that allowed to attribute the language of the text in the automatic mode were not established. This study aims to expand the method for language attribution based on the assessment of the text *irregularity* and define the criteria for the automated text language attribution.

3 Language Attribution

For the experiment, only one font (Times New Roman) is commonly used in electronic documents and ten European alphabetically based languages (Russian-Rus, English-Eng, French-Fra, Italian-Ita, German-Ger, Czech-Cze, Greek-Gre, Turkey-Tur, Swedish-Swe, and Serbian-Srb) have been selected. Spanish and Italian languages have shown nearly similar *I* factor during the preliminary study, such, only Italian one has been left in the analysis.

As the text modeling a corpus, the excerpt from the famous F. Dostoevsky's "Idiot" has been selected. The volume of the excerpt was about 4000 characters that correspond to a typical text sheet volume.

For each language, the text sample is translated and layout as a text page in the CorelDraw. Then, each sample divided into curves and the perimeters (*P*) and areas (*S*) calculated. Based on the known *P* and *S*, the I factors for each sample calculated.

In order to define criteria for the language attribution, all *I* factors values have been normalized to the English language value as English is the most applicable language in electronic documents.

4 Results and Discussion

The results of *I* factor calculation for ten European languages are shown in Table 2. The languages have been already ranked by an increase of *I* factor. The English normalization coefficient is indicated in the last column.

TABLE II. THE RESULTS OF THE *I* FACTOR CALCULATION FOR THE COMMON TIMES NEW ROMAN FONT AND TEN EUROPEAN LANGUAGES

Language	*Characters in the corpus*	*I facror (Times New Roman)*	*Normalized by English*
Greece	4602	53,79	0,98
Italian	4598	54,75	1,00
Franch	4664	59,55	1,09
English	4249	60,72	1,11
Swedish	4243	62,48	1,14
Czech	3879	62,79	1,15
Russian	4128	63,25	1,16
Turkish	4032	65,77	1,20
Serbian	4003	66,17	1,21
German	4641	66,81	1,22

Table 2 also contains the information on the number of characters in a corpus (sample). As we know, the larger the sample, the more precise calculation, and at about 4000 characters, the *I* factor value becomes more or less stable.

As may be seen from Table 2, the order of languages has not changed. Newbies (Swedish and Serbian) show their similarity to Czech and Turkish respectively. However, the distinguishability of the proposed method is high enough to separate languages.

Table 3 shows the English language normalized *I* factor for the languages from the previous study.

By comparison Table 1 and Table 2, it may be argued that the dependence of the *I* factor on the number of characters in a sample is still traced, however, the normalized values are more sustainable. This supports the idea to use them as a main factor of the language attribution.

Comparing Table 3 and Table 2 one may see that normalized *I* factor values for different fonts do not match. Even more, *I* factor calculated on small sample volumes, also, do not match with 4000-characters ones. This fact should be further investigated.

TABLE III. ENGLISH LANGUAGE NORMALIZED I FACTORS

Language	***Characters in the corpus***	***Normalized I facror (Arial)***	***Normalized I facror (Times New Roman)***
Greece	1145	0,88	0,88
Italian	2168	0,93	0,90
Spanish	1133	0,94	0,91
Franch	1721	0,98	0,96
English	979	1,00	1,00
Czech	947	1,01	1,00
Turkish	1077	1,02	1,02
Russian	1009	1,05	1,10
German	1313	1,09	1,09

5 Conclusion

Modern search engines deal with unmarked text corps that need to be analyzed. The attribute of the language is often one of the most important signs of the text, which determines the algorithm for its further processing. A quick definition of the document language is an important technical issue. The growing number of documents available for analysis (Big Data *etc.*) in the environment of web and corporate networks complicates this task.

This paper summarizes two approaches to the quantitative analysis of textual information. As a quantitative assessment, the so-called *I* factor based on the ideas of fractal geometry is used. Calculation of this factor is possible both for electronic texts and for scanned paper texts. Approaches to solving these problems vary significantly, but there is only one ideology.

As an experimental check, the proven approach of calculating the *I* factor for texts in electronic form, which can be represented as vector objects, was applied. Textual samples represented 10 European languages. Comparison of the results obtained and the results of the previous study confirmed that the minimum sample size to obtain a stable result is an excerpt of at least 4000 characters for each selected language.

The use of the *I* factor normalized for a certain language (for example, English) as an attribute criterion is promising, however, research on a large sample of languages and with variable volumes is required.

Extension of the method to the field of higher automation can be carried out both for the first (vector) approach and the second (raster) approach, however, in the second one, large computational power will be required. We believe that both areas of further research are promising.

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, V. Filimonov “Legibility of textbooks: a literature review”. Procedia - Social and Behavioral Sciences, 2015, Vol.174, 1300–1308.

[5] B. Mandelbrot “Fractal geometry of nature”. Moscow, Institute of computer studies, 2002, 656p.

[6] 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.

[7] 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)

[8] D. A. Тарасов, A. P. Sergeev, A. G. Tyagunov, “Assessment of irregularity 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)

[9] 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)

[10] D.A. Tarasov "A method for language attribution based on assessment of text irregularity".in Mathematical Methods and Computational Techniques in Science and Engineering II, AIP, Vol. 1982, 2018, 020006