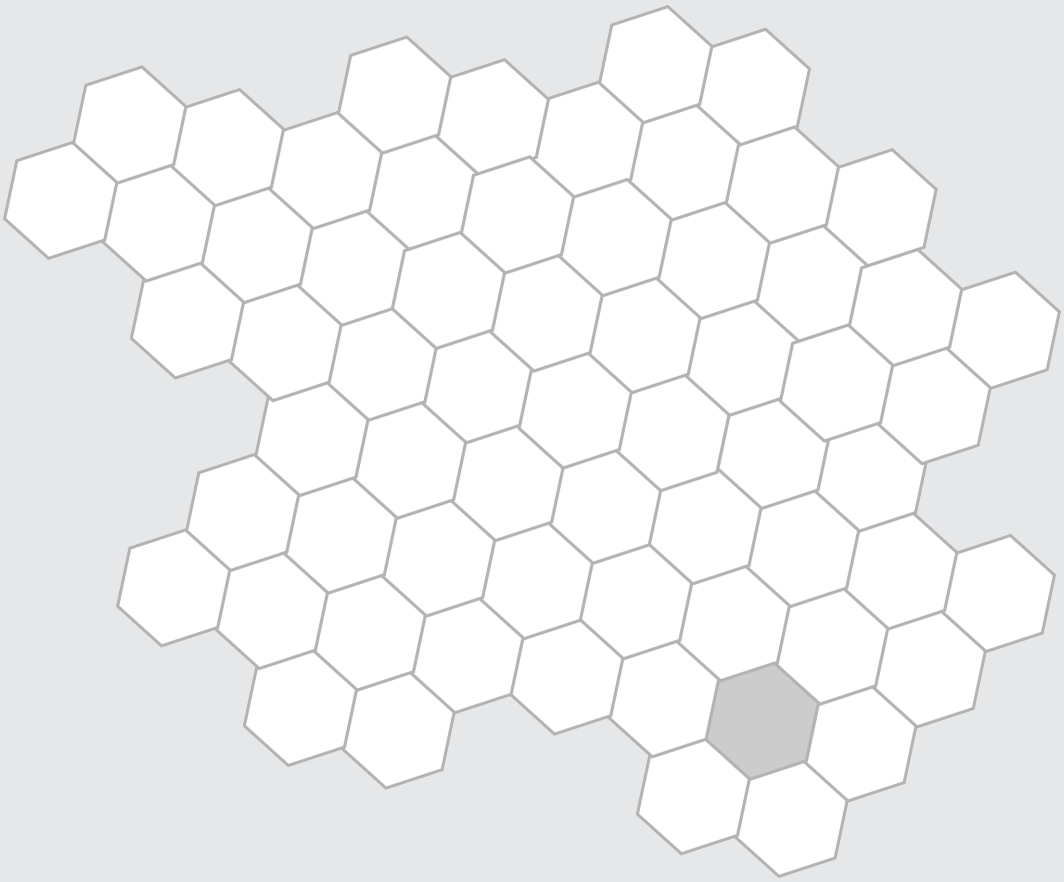
International Journal on

SOCIAL MEDIA

MMM: Monitoring, Measurement, and Mining



I, 2010, 1
ISSN 1804-5251

...OFFPRINT...

International Journal on SOCIAL MEDIA

MMM: Monitoring, Measurement, and Mining

I, 2010, 1

Editor-in-Chief: Jan Žižka

Publisher's website: www.konvoj.cz
E-mail: konvoj@konvoj.cz, SoNet.RC@gmail.com
ISSN 1804-5251

A Look at Wikipedia Readability: Language, Domain and Style

Olga Ogurtsova
Department of French and Romance Philology, Faculty of Philosophy and Arts, Autonomous University of Barcelona, 08193 Bellaterra, Spain
e-mail: assailable@yandex.ru

Mikhail Alexandrov
Department of French and Romance Philology, Faculty of Philosophy and Arts, Autonomous University of Barcelona, 08193 Bellaterra, Spain
e-mail: malexandrov@mail.ru

Xavier Blanco
Department of French and Romance Philology, Faculty of Philosophy and Arts, Autonomous University of Barcelona, 08193 Bellaterra, Spain
e-mail: Xavier.Blanco@uab.cat

Biographical note

Olga Ogurtsova completed a secondary school with a linguistic bias and she graduated from the Russian State Pedagogical University of Herzen in 2008. Her speciality is teaching methods of the Spanish and English languages. At the moment she is a student on a Master's program of the Autonomous University of Barcelona in Spain. The area of her scientific interests is linguistic aspects of internet technologies.

Mikhail Alexandrov is a member of the LexSem Research group of the Autonomous University of Barcelona in Spain. He is a professor of the Academy of National Economy under Russian Government. He is an applied mathematician and author of numerous publications related to mathematical modelling and natural language processing. His current topics of research are machine learning (inductive modelling, clustering) and internet technologies (social networking).

Xavier Blanco is full professor at the Autonomous University of Barcelona (UAB) in Spain. He is author of several large-coverage electronic dictionaries of Spanish for machine translation software and other NLP applications. He is also author of numerous scientific and technical papers concerning translation studies, lexicology, phraseology and lexicography. He coordinates the International Master's program in Natural Language Processing and Human Language Technology in the UAB.

Introduction

State-of-the-art

Wikipedia 'is a multilingual, web-based, free-content encyclopaedia project based on an openly-editable model' (http://en.wikipedia.org/wiki/Wikipedia:About). Wikipedia is written collaboratively by largely anonymous Internet volunteers who write without being paid. Wikipedia content is intended to be factual, notable, verifiable with cited external sources, and neutrally presented. The basic principles of Wikipedia's Arbitration system and committee (known colloquially as 'Arbcom') were developed mostly by Florence Devouard, Fred Bauder and other key early Wikipedians in 2003. The principles can be found in Wikipedia itself (http://en.wikipedia.org/wiki/Wikipedia:About).

Over the years, interest in the Wikipedia phenomenon has been growing. Probably it reached its peak in 2008. At present there is a huge number of works that deal with Wikipedia, and conferences devoted to the Wiki-like resources are held in different cities. There are investigations into the accuracy of Wikipedia, collaborative work of the authors of Wikipedia articles, vandalism in Wikipedia, content of the different language sections of Wikipedia (Giles, 2005; Shapovalov and Malutina, 2009; Belani, 2009; Biuk-Aghai and Lei, 2010), etc.

The homogeneity of Wikipedia style is one the principal requirements for Wikipedia texts. This concerns the homogeneity both of texts that belong to different domains and different languages. Readability is the easiest stylistic characteristic of text to compute, which is why it became the subject of our research.

Readability is defined as the level of complexity of text comprehension, which is determined by certain computable linguistic and stylistic features, such as average lengths of sentences and words, average frequency of rare words and average number of prepositions in phrases, etc. Various indexes are used to evaluate the readability/complexity of a text: the Dale-Chall Readability Formula, Flesch readability index, Farr-Jenkins-Patterson Formula, Fry Readability Formula, Fog Index, Lorge formula, and SMOG Grading (Dale and Chall, 1948; Flesch, 1948; Farr, Jenkins and Patterson, 1951; Fry, 1977; Gunning, 1952; Lorge, 1939; McLaughlin, 1969).

Investigations showed that readability depends on genre (novels, newspapers, scientific papers, etc) and this problem has been considered in publications related to the mentioned indexes. The readability of textbooks for schoolchildren and students are the most referenced topic in index descriptions and their applications (`http://en.wikipedia.org/wiki/Flesch-Kincaid_readability_test`).

The most popular readability index among western researchers is the Flesch index. It uses average sentence length measured in words and average word length measured in syllables: the bigger this index, the higher the ease of reading. There are programs designed for calculation of this index, for example Word Counter for Macintosh OS X or INFLESZ for Windows 9x. y NT/XP (`http://www.legibilidad.com/home/acercade.html`). Some Internet applications contain functions that permit calculation the Flesch index. For example, the Flesch index for different languages can be calculated on the web-page (`http://www.standards-schmandards.com/exhibits/rix/`). There is a program for calculation of the Flesch index for Russian and English text (Oborneva, 2005) but this program is not a free-share one.

The Mikk index is known among Russian researchers (Tuldava, 1975; Makagonov 1998). Mikk's formula includes the same variables as Flesch's formula, namely the average sentence length and the average word length in a text. However, this index reflects the complexity and not the readability of a text. In other words the higher the index is the more difficult a given text is to read. The bibliography lacks references to software that counts the Mikk index. We used the program TextComplexity developed in the department of French philology of the AUB. This program was used also by one of the authors when she was working on her Master's thesis (Ogurtsova, 2010).

It is worth mentioning that we did not find publications where the problem of stylistic homogeneity of Wikipedia texts had been considered. The proximity

of the Wikipedia style to scientific style has not been evaluated as well.

Problem settings

There are two problems to be considered in this article.

1. We want to check whether the administrators of Wikipedia adhere to equal readability requirements for texts in different languages and domains. We have chosen two domains – physics and linguistics, which contrast as natural and humanitarian disciplines. That is why a comparison of texts from these domains would reflect the situation in other domains less contrasted in their content. As concerns the languages, we have chosen English and Spanish, which do not belong to the same language group (as Romance, Slavic or Finno-Ugric languages). The Flesch and Mikk indexes are used to compare texts. The problem consists in the calculation of the mean value of the indexes for groups of texts and testing a hypothesis about the statistical significance or non-significance of the differences.
2. We want to know to which sub-style of the scientific style (properly scientific, popular or didactic) texts from Wikipedia belong. English texts on linguistics representing the three mentioned sub-styles were selected in order to be compared with texts from Wikipedia. In this case we use the Mikk index only for the comparison. As well as in the previous case we need to find the mean of the index for each group of texts and then to test the statistical significance or non-significance of the differences.

The paper contains 4 sections. The next section describes the method of investigation. This method consists of testing the hypothesis about non-significant differences in the readability indexes mentioned above. Section 3 presents the results of experiments. The discussion is included in section 4.

Decision making

Modified Flesch and Mikk indexes

Main formulae

The following Flesch formula is accepted for the English language (Flesch, 1948):

$$\mathrm{IF} = 206.84\text{-}84.6S - -1,015M$$

where S is the average length of a word in syllables and M is the average length of a sentence in words.

The following Flesch formula is accepted for the Spanish language (`http://www.legibilidad.com/home/acercade.html`):

$$\mathrm{IF} = 206.84\text{-}62.3S\text{-}M.$$

Table 1 demonstrates the correspondence between the Flesch index and level of readability

This is the Mikk formula:

$$IM = SLn(M)$$

where S is the average length of a word in syllables and M is the average length of a sentence in words.

The equation of regression for word length in characters and syllables

Revealing syllables in words is a procedure that needs to take into account the linguistic properties of a given language. Unfortunately we did not find in the literature any universal free-share software that could do it. In this situation it is reasonable to consider the possibility to substitute syllables with characters. But such a substitution needs justification.

For this we selected pieces of text from several English documents (newspaper) and took the 100 most frequent words. Then we calculated the number of characters and syllables in each word and constructed a regression

$$y = 2.86x \tag{1}$$

where y is the quantity of characters and x is the quantity of syllables in a word. The coefficient of correlation between the mentioned values is equal to 97,4%.

The same experiment was done with words from the Spanish language. The following dependence was revealed:

$$y = 2.35x. \tag{2}$$

Here the coefficient of correlation is equal to 98,8%.

All the computations were completed with the MegaStat package.

Modification of the Flesch and Mikk formulae

The dependences (1) and (2) obtained were utilised to substitute the number of syllables with the number of characters in the Flesch formulae. Thereby, after substitution the following Flesch formulae were obtained.

For the English language:

$$[\mathrm{IF} = 206.84 - -84.6/2.86N - -1.015M.$$

Here, N is the average length of a word in characters and M is the average length of a sentence in words.

For the Spanish language:

Table 1: Flesch values and level of readability

Flesch index	Readability
70-80	very easy (novels)
60-65	normative (newspapers)
50-55	intellectual level (business editions, literary magazines)
30 and lower	scientific level (professional and scientific literature).

$$\mathrm{IF} = 206.84\text{-}62.3/2.35N\text{-}M.$$

The Mikk formula was modified formally; we only substituted the number of syllables with the number of characters without any transformation. This substitution is reasonable because it changes all the values of tests in the same proportion. So this formal modification has no an impact on the testing hypothesis. Therefore we have the following Mikk formula:

$$IM = SLn(M)$$

where N is the average length of a word in characters and M is the average length of a sentence in words.

One should note that such a substitution of syllables with characters is justified when we measure the average length of words in a whole text. It does not fit for cases when we have to analyse concrete words.

Comparison of indexes

In all our experiments we compare readability of two document sets using Flesch and/or Mikk indexes. If a given index (Flesch or Mikk) has close values for each set then one can say that these sets have close styles from the point of view of readability. The comparison of indexes is performed statistically in the framework of testing the hypothesis about non-significance in differences of index means.

For testing the hypothesis we use the standard technique of p-value (Cramer, 1999). It consists of two steps:

1. One calculates the means (m_1,m_2) and deviations of the means (s_1,s_2) of a given index for two data sets and then forms the so-called t-statistics

$$ts = |m2 - m1|/\sqrt{({s_1}^2 + {s_2}^2)}$$

2. One calculates the probability of the extreme case that random t-statistics reaches this value. $p = P(t > t_s)$

   The lower the p-value, the less likely the result is if the hypothesis is true. Let we fix a level of significance α (10%, 5%, 1%). This level defines the probability of error when we reject the true hypothesis. The technique of hypothesis testing consists in the following rule:

   Hypothesis is accepted if $p > \alpha$

   Hypothesis is rejected if $p \leq \alpha$

   When we reject the hypothesis we can make an error with the probability α (type 1 error)

There are standard functions in all popular packages related to experimental data processing, which calculate p-value for given t-statistics or for two given data sets. For example such functions are included in the list of standard Excel-functions.

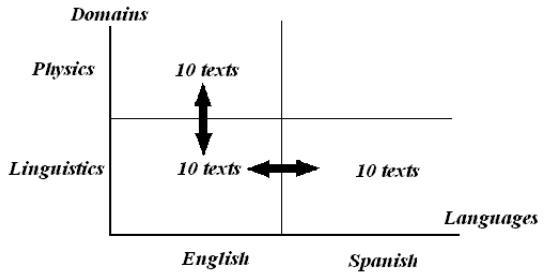


Figure 1:
Languages and domains used in the experiments

Note: If the number of data in each document set is more than 30 then one should use functions for working with the normal distribution. If the number of data is equal to or less than 30 then one should use functions for work with the Student distribution. In the latter case it is necessary to take into account the so-called degree of freedom for the Student distribution. In our case this value is equal to $k = 2n - 2$, where n is the number of documents in each document set.

Experiments

Analysis of the homogeneity of Wikipedia by language and domain

Plan of the experiments

In the first series of experiments we compared different texts from Wikipedia. In order to check the homogeneity of Wikipedia by language we need to take texts from the same domain written in two different languages. In order to check the homogeneity of Wikipedia by domain we need to take texts from two different domains in the same language. In such a way we can essentially reduce the number of experiments and simplify the interpretation of results.

We consider:

- contrasting domains, i.e. linguistics and physics;
- languages from different groups, i.e. English and Spanish.

The plan of the experiments in the co-ordinates language-domain is represented in Figure 1.

The Flesch and Mikk indexes for English and Spanish texts

According to the plan presented in Figure 1 we examined 20 texts on linguistics from the English and Spanish versions of Wikipedia. The average text length was approximately 1,300 words both for English and Spanish documents. The means and variations of the Flesch and Mikk indexes were calculated for each set. Table 2 contains the results of the calculations.

Our goal is to test the null-hypothesis about non-significance of differences in the means for each index. We use here the standard procedure described in section 2.2. The source data is given in Table 2. The degree of freedom is equal to

Table 2: Means and deviations for Flesch and Mikk indexes

Value	Flesch English	Flesch Spanish	Mikk English	Mikk Spanish
Mean	23.62	39.24	16.72	17.34
Deviation	7.59	7.71	1.20	1.29

Table 3: Verification of the null-hypothesis about the non-significance of differences

p-value	p-critical	Result	Flesch index
0.0002	1%	The difference is significant	Mikk index
0.28	1%	The difference is non-significant	

$k = 2n - 2 = 18$, importance level αy = 0.01. Table 3 contains the p-value for each test and the result.

Therefore the Flesch index shows the significant difference for texts in English and Spanish, while the Mikk index does not detect this difference. This fact can be explained by the influence of natural differences in the two languages (English words are shorter and English sentences are longer then Spanish ones) which is not considered by the Mikk index.

The Flesch and Mikk indexes for texts on linguistics and physics

According to the plan presented in Figure 1 we examined 20 English texts on linguistics and physics. The average text length was equal to approximately 1,300 words for documents on linguistics and 1,200 words for documents on physics. The means and variations of the Flesch and Mikk indexes were calculated for each set. Table 4 contains the results of the calculations.

Table 4: Means and deviations for Flesch and Mikk indexes

Value	Flesch Linguistics	Flesch Physics	Mikk Linguistics	Mikk Physics
Mean	23.62	33.41	16.72	15.68
Deviation	7.59	6.36	1.20	0.75

To test the null-hypothesis we complete the same procedure described in the previous item. Table 5 contains the p-value for each test and the result.

Both Flesch and Mikk indexes revealed an absence of significant differences between the texts on physics and linguistics.

The graphical illustration of the results obtained relative to the average values is presented in figures 2 and 3.

Scientific style of Wikipedia

Functional styles

Functional style is a variety of the literary language performing a specified function in communication (Solganik, 1997). In this paper we examine the scientific style and its sub-styles: the properly scientific, popular and didactic.

Table 5: Verification of the null-hypothesis about the non-significance of differences

p-value	p-critical	Result	Flesch index
0.01	1%	The difference is non-significant	Mikk index
0.03	1%	The difference is non-significant	

Figure 2: Flesch indexes for the texts

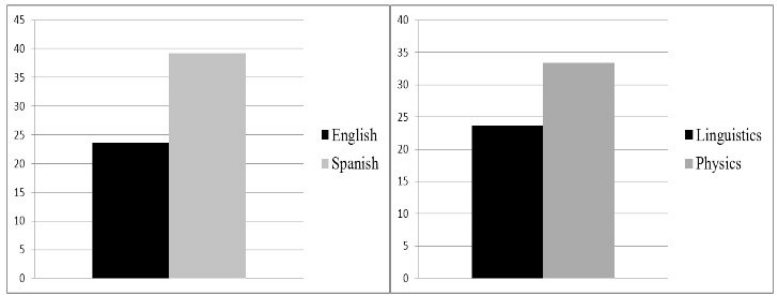
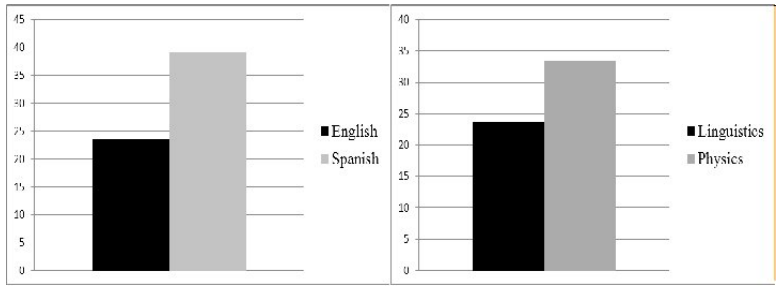


Figure 3: Mikk indexes for the texts



Scientific style is characterised by the logical sequence of phrases and it is characterised by accuracy, brevity and absence of ambiguity. The research paper is forwarded to a reader specialising in a concrete scientific branch and possessing knowledge at about the same level as the author of article, the sender.

The specificity is inherent also in the popular sub-style. The receiver of such articles is a person interested in this or that science. Within the limits of this sub-style some deviations from norms are supposed – the use of words in their figurative sense is possible.

Documents with a didactic sub-style are addressed to future specialists, students and schoolchildren. Its purpose is to teach and to describe the facts that are necessary for the acquisition of material.

We consider the degree of correspondence of Wikipedia documents to the above mentioned sub-styles of the scientific style from the point of view of text complexity. Text complexity is evaluated here with the Mikk index.

Certainly, scientific style and its sub-styles are not restricted only by text complexity. The full characteristic of the style requires testing other formally-statistical and informally-linguistic indicators. For example, harmony of texts and richness of text vocabulary refer to the first one, while specific idioms refer to the second. However, in the present work we limit our consideration to text complexity only.

In our experiments we used documents on linguistics in English. We selected 10 texts from Wikipedia, 10 scientific articles, 10 popular articles and 10 manuals. Therefore we had 40 texts in total. The sources of the scientific articles are various theses on linguistics and research described in articles for scientific journals. The popular texts were taken from electronic newspapers or journals for a wide range of readers and from encyclopaedia articles. The didactic materials are more varied. We were working with fragments from textbooks on linguistics for high school students and for students of the first years of university (both from faculties of languages and from technical faculties).

The means and deviations of the Mikk index were calculated for each group of texts. The results are presented in Table 6.

Table 6: Values of the Mikk index for texts on linguistics

Category	Mean	Variation
Wikipedia	16.72	1.2
Scientific articles	17.04	1.1
Popular articles	15.88	1.35
Didactic materials	12.67	2.44

The means of the Mikk index are presented on the Figure 4.

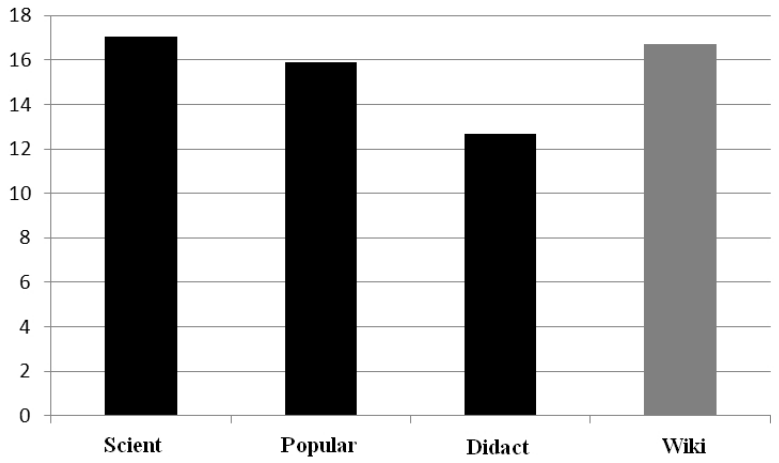
We tested the significance of differences in Mikk index between the Wikipedia articles and the other three groups of texts. We verified the null-hypothesis about the non-significance of differences of the means. The verification of the hypothesis was done as described in part 'Comparison of indexes'. The results are presented in Table 7.

Table 7: Verification of the zero-hypothesis about the non-significance of differences

	p-value	p-critical	Result
Scientific articles	0.54	1%	The difference is non-significant
Popular articles	0.16	1%	The difference is non-significant
Didactic materials	0.0004	1%	The difference is significant

Therefore the level of complexity of Wikipedia texts differs significantly from the level of complexity of didactic texts. It is close to the level of complexity of scientific and popular texts.

Figure 4: Values of the Mikk index for different text categories



Conclusion

Discussion

- We revealed the dependences between average number of characters and syllables in English and Spanish words. This relation has a high correlation (97%-99%). Based on this relation we could modify the Flesch formula and justify the possibility of using the Mikk formula with a formal substitution.
- A significant difference in the Flesh index for English and Spanish texts was shown, although the Mikk index did not detect such a difference. In the experiments, documents on linguistics were used.
- It was shown that there were no significant differences in Flesh and Mikk indexes for linguistics and physics. In the experiments, English documents were used.
- The level of text complexity for Wikipedia articles is close to the text complexity of scientific and popular documents and differs only from manuals. In the experiments, documents on linguistics were used.

Our conclusions were based on very limited document sets. So, the obtained results can be considered only as preliminary ones, which should be tested once more on a larger corpus of documents.

Future work

In the future, we consider:

- repeating the completed experiments on large data sets containing dozens and even hundreds of documents;

- enlarging the number of languages, in particular, to consider texts in the Russian and German languages;
- enlarging the number of domains, in particular, to consider economics and history;
- considering other formally-statistical indicators of style, such as the degree of lexical richness of text.

References

Wikipedia:About [on-line]. (c2010): [cit. 2010]. Available at: `http://en.wikipedia.org/wiki/Wikipedia:About`.

Belani, A. (2009): *Vandalism Detection in Wikipedia: a Bag-of-Words Classifier Approach Master's*. Cornell University„ .

Biuk-Aghai, R. P.; Lei, Keng Hong (2010): Chatting in the Wiki: Synchronous-Asynchronous Integration.. In: *Proceedings of the 6th International Symposium on Wikis and Open Collaboration*.. Gdańsk (Poland) .

Cramer, H. (1999): *Mathematical methods of statistics*. Princeton University Press„ .

Dale, E., Chall, J. S. (1948): A formula for predicting readability.. *Educational research bulletin*, Jan. 21 and Feb. 17, 27, pp. 1–20, pp. 37–54.

Flesch, R. (1948): A new readability yardstick.. *Journal of Applied Psychology*, Vol. 32, pp. 221–233.

Farr, J. N., Jenkins J. J., Paterson D. G. (1951): Simplification of the Flesch Reading Ease Formula.. *Journal of Applied Psychology*, Vol. 35, No. 5, pp. 333–357.

Flesch-Kincaid readability test [on-line]. (c2010): [cit. 2010]. Available at: `http://en.wikipedia.org/wiki/Flesch-Kincaid_readability_test`.

Fry, E. (1977): *Elementary Reading Instruction*. New York, .

Giles, J. (2005): Internet encyclopedias go head to head.. *Nature*, December 15, pp. 900–901.

Gunning, R. (1952): *The technique of clear writing*. New York : McGraw-Hill, .

...la legibilidad? (in Spanish) [on-line]. (c2010): [cit. 2010]. Available at: `http://www.legibilidad.com/home/acercade.html`.

Lorge, I. (1939): Predicting reading difficulty of selections for children.. *Elementary English Review*, 16, pp. 229–233.

Machine Learning [on-line]. (c2010): [cit. 2010]. Available at: `www.machinelearning.ru`.

Makagonov, P., Alexandrov, M. (1998): Analysis of contents of educational courses with special statistical and graphical methods.. In: *Proc. XV World Computer Congress, Conf. "Téleteaching '98"*.. Vienna : Austrian Comp. Soc., pp. 679–683.

McLaughlin, G. Harry (1969): SMOG Grading – a New Readability Formula.. *Journal of Reading*, Vol. 12, No. 8, pp. 639–646.

Oborneva, I. V. (2005): Mathematical model for evaluation of didactic texts.. *Proc. of Moscow State Pedag. Univ., series 'Informatics'*, Vol. 4, No. 1, pp. 141–147.

Ogurtsova, O. (2010): The statistical index of readability as a formal indicator of a style of scientific/popular/didactic texts. .

Readability index calculator [on-line]. (c2010): [cit. 2010]. Available at: `http://www.standards-schmandards.com/exhibits/rix/`.

Shapovalov, R., Malutina, A. (2009): *About the mission of language sections of Wikipedia different from the English Wiki-conference 2009*. Saint-Petersburg, .

Solganik, G. Y. (1997): *Stylistic of text: Manual*. Moscow, .

Tuldava, Yu (1975): About Measurement of Text Difficulties.. In: *Proc. Of Tartu State University*.. , pp. 102–120.

The authors thank Prof. Julio Murillo from the Department of French and Romance Philology of the Autonomous University of Barcelona for his valuable advice and consultations.

Call for Papers

Papers to be included in the next issue should be preferably focused on topics related to social-networks in one or more of the following subjects (the list is indicative rather than exhaustive):

Sentiment/Opinion Analysis in Natural-Language Text Documents

Algorithms, Methods, and Technologies for Building and Analysing Social Networks

Applications in the Area of Social Activities

Knowledge Mining and Discovery in Natural Languages Used in Social Networks

Medical, Economic, and Environmental Applications in Social Networks

Submitted papers should not have been previously published nor be currently under consideration for publication elsewhere. Each of the submitted research papers should not exceed 26 pages. All papers are refereed through a peer review process.

Submissions should be send in the PDF form via email to the following address: `SoNet.RC@gmail.com`

Accepted papers are to be prepared according to the instructions available at `http://www.konvoj.cz/journals/mmm/`.