International Journal of Computer (IJC)
ISSN 2307-4523 (Print & Online)

http://ijcjournal.org/

Software Construction for the Estimation of the Linguistic Level and Test Difficulty

Apostolos C. Klonis*

Informatics Secondary School Teacher, M.Sc., PhD., Lefkonas Serron, Serres 62100, Greece
Email: apoklonis@gmail.com
Email: atklonis@itl.auth.gr

Abstract

For this survey a new linguistic level evaluation and test measurement software has been created. This particular software has assisted in detection matters regarding readability and it has also allowed text readability measurement with the use of common grading systems, including readability measurement formulas. This system accepts various examination topics, which are classified according to the level of difficulty and where all kinds of tests are represented and it controls all the linguistic level and difficulty goals. The choice of topics and its inclusion is conducted with the sampling method. During this experimental application of our software, a field survey was conducted during which not only university students but also a lot of internet users were called to evaluate this programme.

Keywords: Linguistic level software; test; linguistic level; text; readability rates.

1. Introduction

The new software which has been developed is a more reliable and valid solution for the evaluation of difficulty of the Italian language tests, which are used in the development of exam topics for the various levels of linguistic certification. During this experimental application of our software, a field survey was conducted during which not only university students but also a lot of internet users were called to evaluate this programme. This evaluation was conducted in two stages. The first was right after the software completion and the second one after its evaluation and its possible improvements. Also, further user characteristics have been examined, since the programme effectiveness may be influenced by other variables such as the age, the gender, the training and in general the educational level of the users. There were also closed-format questions which made the processing and the analysis of the collected data easier.

---

* Corresponding author.

Using statistic techniques and methods appropriate for categorical data [10], it was easy to study the relations among the variables. For this purpose the theories of Factor Correspondence Analysis and of Multiple Factor Correspondence Analysis have been used. All the above have been critically compared in theory and in practice, so that the interested readers can understand these methods and can get as much information as possible that have helped them in this application.

More specifically, the internet software that has been created for this purpose has the following functions:

- The readers can insert their text and the software can extract the results and the readability grades as well as the test difficulty level
- The user can see some further system functions, such as word, syllable, character, sentence counters, characters per word, syllables per word, words per sentence.
- Also, there is the possibility of text readability in a whole website, with the user typing the corresponding web address.
- And, finally there is the possibility to count the readability of .doc, .docx, and .pdf files, as long as these files are uploaded in the software.

During the research, there have been particular efforts, so that the writing and listening comprehension test difficulty criteria could be gathered, as well as the writing and oral production test difficulty criteria for the Italian language. Since they will be inserted in the software after the processing of these linguistic tests, these texts can be classified in the corresponding levels and in the corresponding test difficulty levels. Finally, the reader will possess a readability software which: α) uses more parameters with increasing discreet power among the linguistic level, b)can give direct correspondence between readability and linguistic level and c) estimates the possibilities which correspond to all linguistic levels and their degree of difficulty.

2. Stylometric features of text readability

The text reading degree of difficulty is interwoven with the symbol notation, the decoding of the content of words, the identification of symbols with appropriate words that these symbols denote, with the ultimate goal being the penetration in meaning of the text. Only then can the text function as the basic element of learning and of mental cultivation for the student. Of course, in reading comprehension a lot of factors intervene, which are related to the reader, to the text itself as well as to the environment where the reading procedure takes place. The difficulty of the text reading depends on factors which are formed by the text author, such as: the sentence length, the word length, the paragraph length and in general, the text formatting. Furthermore, the linguistic stylometric features have been counted and are numerous, which shows that they can be exploited in the estimation of the reading difficulty and of the degree of difficulty of a test [15]. Two of the most well-known stylometric features, which are the average word length and the average sentence length, are the basics of the most well-known readability formulas. The most important stylometric features are analysed below [16].

- Words: The total length of a text in words
- Type/token ratio (TTR): The ratio of the number of vocabulary units (types) to the number of words

(tokens) of the text. The longer the ratio, the "richer" the vocabulary of the text.

- AWL (Average Word Length): The average length of words of each text is estimated with the character as the basic unit of measurement.
- WLsd (Word Length standard deviation): The standard deviation of the average length of the words in the text
- Sentence length: The average length of sentences in a text counted in words.
- SLsd (Sentence Length standard deviation): The standard deviation of the average length of sentences of each text.
- Perc_HapL: The ratio of words which appear in the text once is estimated.
- Perc_DisL: The ratio of words which appear in the text twice is estimated.
- Dis_HapL: This ratio has been suggested as indicative of the writing style.
- LD (Lexical Density): The ratio of the percentage of content words to the percentage of function words.
- Entropy: The entropy of each text is estimated, which is the degree of the organization and predictability of lexical frequencies.
- Relative Entropy: The ratio of the theoretically maximum entropy of a text to the observed entropy. A text presents maximum entropy when every word it would include would appear only once and so, all the words in this text would be "HAPL". The bigger the relative entropy, the less standardized the text and so "richer" in vocabulary [8].

To investigate the usefulness of stylometric variables in the categorical evaluation of the text difficulty [14] and of the degree of difficulty of a test, the statistical method of logistic regression is used. The logistic regression [13] in its simplest form, as well as in the simple linear regression, is a statistical model which uses one dependent and one independent variable and produces a linear equation which includes one invariant (b0) and the regression quotient (b1) for the independent variable (x). This linear equation equals the natural logarithm of the supplementary odds of the fact of the dependent variable [2].

3. The Analysis of Factors which affect text readability

The analysis and study of the factors which affect readability have been the object of study of many linguistic and psychological studies after 1950 and have led to the adoption of some direct or indirect indicators, which count the easiness or difficulty of reading and comprehension, regarding the content of a text by an average student. One such indicator is the readability easiness degree or indicator [3]. With the term readability or readability easiness we mean the easy, fast and correct recognition of the forms of letters, words and symbols of a text, which entails its comprehension and decoding. The special features of each text have led a lot of linguists to the formulation of mathematic formulas which connect them and make up an objective way of estimation of text readability. For several years there have been various formulas of text difficulty measurement [6] like formulas which use traditional readability criteria, formulas which are based on cognitive theories, formulas which are based on analytics, as well as on mixed methods. Below are the most known formulas of text readability estimation. There are 3 methods of the objective estimation of readability:

1. Question and Answer Technique [7] in order to carry out this technique, students of different ages read a text. After that, questions are posed so that the comprehension level can be defined and by that the reading age of the students is estimated. Nevertheless, this is difficult to be done by teachers [1].
2. Text comparison with a particular list of words [9] the percentage of words which are not included in this list is counted [5], as well as the reading age of the students, so that the readability degree of the text can be estimated. Based on this technique and well-known in the Anglo – Saxon countries are the Dale - Chall tests [11].
3. According to the sentence length and the number of syllables: these are objective measurements, which are widely used [4] They are expressed as mathematic formulas (or charts) which are based on a big quantity of research data. The readability also shows the reading level of the text, manifesting the age of the average reader – student, who can easily comprehend it [12].

4. The creation and design of the linguistic software

For the creation of the graphic environment of Calculator.exe the programme Qt Creator has been used. The Qt Creator is an embedded development environment of C++, Javascript and QML, which is part of the SDK for the framework of the application development of Qt GUI. It includes an optic tool of error detection, an embedded pattern GUI and a form designer. The features of the processor include the structure signaling and the automatic completion. The creator Qt uses the compiler C++ from the completion of GNU Compiler in Linux and FreeBSD. In Windows it can use the MinGW with the default setup and it can use the Microsoft Debugger Console when it is compiled by the source code. The Clang is also supported. The creator Qt includes a project administrator which can use a variety of project forms, like .pro, CMake, Autotools and more. A project file can include information, such as which files are included in the project, the steps of adapted creation and the settings for the execution of the applications. Also, it includes a programme of code processing and embeds the Qt Designer for the design and the construction of graphic interactions between the user (GUI) and the widgets Qt . The code processor in the Qt Creator supports the structure signaling for different languages. Besides that, the code processor can analyze the code in languages C++ and QML and have as a result the code completion and the help of a semantic browsing sensitive to the environment. It is possible to compose and adapt the graphic elements or the dialogues and to try them using different styles directly to the processor. The graphic elements and the forms which are created in combination with the Qt are embedded in a programmed code, using the signals QT and the reception mechanism. Finally, it provides support for the development and execution of the Qt applications for desktop environments (Windows, Linus, Free BSD and Mac OS), mobile devices (Android, BlackBerry, iOS, Maemo and MeeGo) and embedded Linux devices. The creation settings allow the user to change creation targets, various versions of Qt and creation formations. For the targets in mobile devices, the Qt Creator can create a set up package, install it in a mobile device which is connected to the development computer and install it there. The setup packages can be published in Ovistore. The software calculator.exe (text readability tool) is destined for educational use by people of different ages who are familiar with the use of the computer and have as a goal the overview, analysis and evaluation of the information exerted by the software, such as the estimation of the readability degree of an Italian text, the test level of difficulty, the periods, the amount of words and a lot more information which we will deal with below. For the requirements of the system a computer of at least 2.53 GHZ is required, if it is a personal computer, and a memory of 2 GB RAM. The

software home screen provides access to the different functions of the applications for the creation of reports, after the insertion of Italian texts by the final user. The user can select the text analysis and thus a new window with text information will be opened. Automatically, all the feature values will be altered and the final readability degree (calculator) as well as the corresponding number of texts will be estimated. Also, the test level of difficulty in the corresponding level will appear and will be illustrated with the corresponding chart. The second option has to do with the printed form of the software, since all the test information is printed in a list by our computer printer. In the third option the information appears in the form of tabs separated in categories and subcategories and finally the last option is the chart. Everything here is illustrated with different charts and the user is able to see them with schematic illustration.

5. The software communication graphic environment calculator.exe

The software condition icons are the following:

1. Application termination: this is how we terminate the application.
2. File Opening: the file opens locally by our computer. Any document can be inserted as long as it doesn't contain any formatting. We can also choose any text from the internet and with the "Copy – Paste" method we can insert it in our software.
3. Text Analysis: selecting the "Text Analysis" button, we will observe that the information we need has been extracted. The text readability degree will be estimated with the corresponding comment in the next column. Also, the cognitive level with the subsequent characterization in a chart form will be estimated.
4. Projection in Printable html: here there is a detailed text analysis and test difficulty measurement in Printable form.
5. Projection in Tabbed html: at this stage, there is a detailed text analysis and test difficulty measurement in the form of tabs.
6. Chart: it is a chart of all the data.
7. Software help icon: several information about the calculator.exe software.

Moreover, in the graphic environment of the programme there are various tags, such as:

- Towards Analysis Text: in this tab, we insert the Italian text to see the text level of difficulty and the degree of difficulty of the text. After the insertion of the text we press the analysis button.
- Sentence periods: the sentence periods are separated until the punctuation marks and definitely until the full stop.
- Number of sentences: the software is programmed to count the numbers of sentences until the full stop.
- Words – Characters: in this tab and after the analysis of the text, the words which are in the dictionary appear. According to which words appear more often, we get the corresponding level of difficulty. Also, the characters which are in the text are counted. We should consider that the same characters are not taken into account. The combination of words which are in the dictionary and the number of characters give us the degree of text difficulty.

- Tokenization: here we have a more detailed analysis of the inserted text to find its degree of difficulty. This index is important for the test result. Thus, not only the sentence separation takes place, but also the grouping of words, letters and most frequently used words in an Italian text.
- Finally, the respective text features are observed such as the modified Calculator index, the number of periods, the number of sentences, the number of words, the number of characters, the words per sentence, the sentences per period, the characters per word, the words in the dictionary and the words off the dictionary with their respective values in the next column.

At the execution of the calculator.exe software, after the insertion of the Italian text, some of its features are modified such as:

- Calculator: this is our index of the Italian language, modified in our software. After the text is inserted, it returns the value which corresponds to the degree of difficulty of the test and to the level of difficulty.
- Number of periods: the sentence periods are separated until the punctuation marks and definitely until the full stop.
- Number of sentences: the number of sentences always starts from the capital letter and ends at the full stop.
- Number of words: this counter counts the words which are in the dictionary of the A1, A2, B1, B2, C1 levels.
- Number of characters: the characters which are in a text
- Words per sentence: the number of words in a sentence, given that one sentence has an average of 10 words.
- Sentences per period: in this tab we've got the total amount of sentences in a period. It is estimated by the number of sentences divided by the number of words.
- Characters per word: this counter counts the characters which correspond to a word. To demonstrate this, we get an average of 100 words. Punctuation marks, quotation marks and spaces are not taken into consideration.
- Words off dictionaries: These are the words that are not in the dictionary. These words are considered unknown and are not taken into account. So, if the user types anything and in a different language, this will not be considered for the text comprehension and for the test level of difficulty.

Acknowledgments

The authors thank teachers who give knowledge about how to read papers and write papers.

6. Conclusions

The software calculator.exe is the only one which will be released in the market with the parameters mentioned above and will measure the degree of difficulty of the tests that will be inserted and the level of difficulty of the text readability in the Italian language. It is easy to use with amazing results. Finally, it will positively contribute

in the field of education, since it will be widely used by Universities and by the exam committees of the State Certificate of Languages.

References

[1]. Ajina, A., Laouiti, M., & Msolli, B. (2016). Guiding through the Fog: Does annual report readability reveal earnings management? Research in International Business and Finance.

[2]. Bengio, Y., Ducharme, R., Vincent, P., Janvin, C., De Mulder, W., Bethard, S., … Mikolov, T. (2015). Statistical Language Models Based on Neural Networks. Computer Speech & Language.

[3]. Bormuth, J. R. (2006). Readability: A New Approach. Reading Research Quarterly.

[4]. Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery.

[5]. Catanzaro, B., Sundaram, N., & Keutzer, K. (2008). Fast support vector machine training and classification on graphics processors.

[6]. Caylor, J. S., Sticht, T. G., Fox, L. C., & Ford, J. P. (1973). Methodologies for Determining Chapelle Reading Requirements of Military Occupational Specialties. In Human Resources Research Organization, Alexandria, VA.

[7]. Chapelle, O., Haffner, P., & Vapnik, V. N. (1999). Support vector machines for histogram-based image classification. IEEE Transactions on Neural Networks.

[8]. Chih-Wei Hsu, Chih-Chung Chang, C.-J. L. (2008). A Practical Guide to Support Vector Classification. In BJU international.

[9]. Contreras, A., García-Alonso, R., Echenique, M., & Daye-Contreras, F. (1999). The SOL formulas for converting SMOG readability scores between health education materials written in Spanish, English, and French. Journal of Health Communication.

[10]. Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. Machine Learning.

[11]. Dale, E., & Chall, J. S. (1948). A formula for predicting readability: Instructions. Educational Research Bulletin.

[12]. Dale, E., & Chall, J. S. (1949). The concept of readability. Elementary English.

[13]. Daniela D, M. V., & Maria Celeste Pirozzoli, A. (2013). Application of a Readability Score in Informed Consent forms for Clinical Studies. Journal of Clinical Research & Bioethics.

[14]. Mikros, G. K. (2013). Authorship Attribution and Gender Identification in Greek Blogs. Methods and Applications of Quantitative Linguistics.

[15]. Mikros, G. K., & Argiri, E. K. (2007). Investigating topic influence in authorship attribution. CEUR Workshop Proceedings.

[16]. Mikros, George, & Perifanos, K. (2015). Authorship Attribution in Greek Tweets Using Author's Multilevel N-Gram Profiles. American Printer.