International Journal of Science Academic Research
Vol. 05, Issue 07, pp.7860-7864, July, 2024
Available online at http://www.scienceijsar.com

ISSN: 2582-6425

Research Article

HOW VOCABULARY AND GRAMMAR INFLUENCE THE READABILITY AND LANGUAGE LEVEL IN WRITTEN PRODUCTION

***Chrysovalantou Kapeta**

Department of Italian Language and Literature, Aristotle University of Thessaloniki, 54124, Greece

Received 12th May 2024; **Accepted** 15th June 2024; **Published online** 30th July 2024

Abstract

Scientists have been exploring the field of Readability for many decades through various tools, such as special software or indicators (Gulpease, Dale-Chall, Gunning Fox, etc.), in order to make careful measurements of the Readability degree. The purpose of the present research is to identify those criteria, which first determine the level of difficulty, and second, the language level of a written text. All data were collected from the examinations of the Greek Certificate (KPG) of May 2015 and November 2016. 316 written texts, including both levels, B1 and B2, were digitized manually in Word form. In the second phase, they were measured by using the Read-It tool, and the values and results of this process were evaluated. Through SPSS.24, and in particular, with factor analysis, the final results were achieved. Summarized, inter alia, the following were found: 1. The texts produced by Greek users of the Italian language show that the Readability degree seems to vary by language level and degree of difficulty both through the use of vocabulary as well as grammatical and syntactic features. 2. There seems to be a great difficulty in producing secondary texts, correctly worded in Italian, at both levels. 3. Another phenomenon that reduces both, the readability grade and the languagelevel is the vocabulary confusion with other languages, e.g. we see Greek users writing English or Spanish words confusing them with Italian. These valuable findings and this research are likely to pave the way for future scientists to delve even deeper into the parameters of writing with the ultimate goal of developing more advanced software that will help to improve the use of languages by foreign users and to prepare tests more accurate and fair by certification entities.

Keywords: Validity, Language Level, Readability, Measurement, Vocabulary, Grammar.

INTRODUCTION

The present investigation focuses on written production. Since there has also been a lot of research on this discipline in previous decades (Johansson, 2009), it would be very useful to try to define in what sense a produced text varies from the oral production. Then, finding the most suitable criteria with which a written text is defined as easy or difficult with respect to the language level is of great importance. Written production could be more organized, because while writing, the author is able to look at the text he/she has produced. Reading and writing is a process surrounded by multidimensional factors (Spiro & Taylor, 1980). Precisely these factors have prompted a more intense analysis to find such characteristics that can change the language level and the degree of text difficulty. According to Dell'Orletta, an automatic tool for analyzing the readability of a text for the Italian language is the READ-IT tool (Dell'Orletta, Wieling, Cimino, Venturi & Montemagni, 2014: 164) designed and developed by the Natural Language Processing Laboratory (ItaliaNLP Lab) of the "Antonio Zampolli" Institute of Computational Linguistics (ILC) of the CNR in Pisa/Italy. Through the identification of its areas of complexity, READ-IT was conceived to also provide support for the simplified drafting of a text. This tool implements an "advanced" readability index from a technological point of view based on the multi-level linguistic analysis of the text, conducted by digital means that represent the automatic measurement of the Italian language. This index allows us to calculate the readability of texts whose corpus is composed of others that concern easy or difficult reading according to a particular classification (Tong & Koller, 2001).

Corresponding Author:** ***Chrysovalantou Kapeta,
Department of Italian Language and Literature, Aristotle University of Thessaloniki, 54124, Greece.

This classification is performed by monitoring a series of linguistic characteristics that must be measured automatically and is carried out by a statistical classifier that associates the texts. The named index is a software prototype to access the readability evaluation of a text. READ-IT can analyze texts (an entire document or a sentence) and assigns a score that quantifies readability. In addition, it is a classifier based on Support Vector Machines (SVM)(Gunn, 1998: 1).It is a set of features and a training corpus with which a statistical model is created using the statistical functions extracted from the training corpus in the readability evaluation concerning invisible documents and sentences. There are functions used to build statistics and the model can be parameterized through a file configuration for evaluating the readability of the text (Dell'Orletta, Montemagni & Venturi, 2011). According to Panizza (2016), READ-IT is based on the results of monitoring a series of linguistic characteristics found in a corpus from the output of different levels of linguistic annotation: recording as a lemma, morph-syntactic annotation based on dependencies. Thanks to this monitoring methodology, the linguistic profile of a text is reconstructed on the basis of the distribution of linguistic features that range between different levels of linguistic description: lexical and morph-syntactic elements (Panizza, 2016). In addition to the Gulpease index (Lyding *et al.,* 2014), READ-IT conducts the global evaluation of text readability against four different indices calculated in accordance with four different configurations for text characteristics.

- ✓ BASIC: In this model, the characteristics considered are those used in traditional measures of text readability (i.e. sentence and word length).

- ✓ LEXICAL: This model focuses on the lexical characteristics of the text (i.e. the composition of the vocabulary and its lexical richness).
- ✓ SYNTACTIC: This model is based on grammatical information, in other words on the combination of morph-syntactic features.
- ✓ GLOBAL: It is a model based on the combination of all the traits considered by the other models (Panizza, 2016).

A characterizing feature compared to the international literature on the subject consists in an evaluation of readability divided into two levels: the document and the single sentence. The evaluation with respect to the sentence was explicitly designed to provide support to the editor of the text and guide him in the review and simplification process (Panizza, 2016: 145-146). The starting point of this scientific work would be a test that is recognized as valid and fair for all candidates (Moss, 1994: 9), which is based on objectively suitable, in order to distinguish the language level and the degree of difficulty. Considering the above, we arrive at a better result in producing more understandable texts which can then be evaluated under more scientific parameters, therefore, more effective and productive with regards to language teaching.

If certification is so crucial to the validity (Chan: 2013), professional rehabilitation and development of people in Greece in the tests used for this purpose should be guaranteed to the greatest extent possible:

a) Their relevance for the level of learning of the languages in question;
b) The consistency of their content in the various examination periods;
c) Verification of the level of difficulty regarding the contents.

The absence of the conditions for the effective functioning of the learning test regarding the languages mentioned above can lead to a limitation of the validity concerning the test and therefore of its effect, since the level of linguistic competence is divergent (Venturis, 2018).

There are also following parameters that influence the result of a written production, which matter a lot and are found in other certifications such as CILS (Matthiae, 2010).

- Fluency (good/various breaks/blanks).
- Communicative effectiveness (the message is intelligible/practically incomprehensible/blank paper).
- Morph-syntactic correctness (almost no errors/some errors/many errors).
- Lexical appropriateness (good/acceptable/insufficient).
- Spelling (does not compromise the message/compromises it often/commonly compromises it).

To simplify a text, there are criteria such as those that should be cited by Frigo, Zuppiroli and Pagani (2007: 31).

1. The information is ordered logically and chronologically.
2. The sentences are short (20-25 words).
3. Texts do not exceed 100 words.
4. Coordinated sentences are preferably used.
5. Basic vocabulary is used.
6. Words that are not included in the basic vocabulary are explained.
7. The name is repeated, avoiding synonyms and pronouns.
8. The order S V O (subject, verb, object) is respected in the construction of the sentence.
9. Verbs are used in finite moods and in the active form.
10. Personifications are avoided (e.g. the Senate becomes the Senators).
11. Impersonal forms are not used.
12. Title and images serve to reinforce the understanding of the text.

As regards simplified texts, we observe the following points (Frigo, Zuppiroli & Pagani, 2007).

- Pay attention to the characteristics of the text.
- Also pay attention to images, graphs and references that accompany the text.
- Greater interactivity and expansion of the text.
- Study paths that take into account different phases in language acquisition and integrate linguistic and disciplinary objectives.

It is very stimulating for readers of current research to know that here we have at our disposal productions written by Greek users who are not native Italian speakers. In fact, the teaching of Italian as a second language in Greece is limited. In addition, it must be taken into consideration that students do not have the opportunity to learn Italian in all classes in Greek public schools (Venturis, 2020). In conclusion, the present product aims at a future investigation by upcoming researchers who would like to collect important data and results from similar studies for each foreign language included in the KPG exams to find out the difficulties or ease of Greek users according to the language levels proposed by the Common European Framework for Languages (Mariani: 2014). Furthermore, it would be very interesting to find results from similar exams in other countries where the Italian language is taught to compare the language and readability level in texts produced by Greek and other foreign users, such as Germans for example. The purpose of the present research is to find criteria for determining the language level and the degree of readability based on the use of spelling and vocabulary.

MATERIALS AND METHODS

The main purpose is to find factors with which we could distinguish whether a text is oriented to the B1 or B2 level, always keeping in mind that these are productions written by Greek citizens. In fact, this last particular reflection would be the starting point for having new useful information, provided by the final product of this research that will help foreign language users to improve their language skills. Furthermore, it will help the constructors of different tests to produce exams based on real and scientifically approved factors, and first of all, based on the validity and fairness of the tests created for foreign candidates. Therefore, 316 texts of the KPG exams produced by non-native Greeks should have been digitized. All the texts produced (316 in total) were written with accuracy manually using the software Word (Windows 2010). The Gulpease index and READ-IT tool were used to process all the data of the produced texts. Through the READ-IT and Gulpease indexes, the variables were found, and for the final results, the IBM SPSS STATISTICS VERSION 24 software was chosen.

In the end, all the variables collected from the SPSS table were analyzed, using SPSS graphs and tables to arrive at the conclusions and results of the hypotheses mentioned in the introduction. For the research, we see in table 1 in a descriptive way the most necessary data:

Table 1. Overview of data used

Number of texts chosen	316
Source	Greek State Certification "KPG"
Period	May 2015-November 2016
Language Level	B1, B2
Formulas used	Gulpease, ReadIT
Statistical analysis program	SPSS.24

RESULTS AND DISCUSSION

When there are grammatical or lexical errors, the level of textual difficulty seems to be low (Zawacki& Habib). In the fourth text of this study, there are three unknown and non-Italian words. The words *Mesogia, honora and risponsibile* are elements that change the textual difficulty of the B2 level from a lexical point of view. We would say that errors of this type are probably those that can influence a text produced at any language level, also because such texts including these errors are not comprehensible by Italian native-speakers/ evaluators. The reader may be able to imagine the meaning through the whole context, but this does not help him to understand what the composer is thinking of. Consequently, words that do not exist in Italian are probably not suitable for an intermediate level (B1 and B2). Instead, if we were talking about the elementary level (A1 and A2), it could be less demanding to use a complicated vocabulary and correct grammar, since through the elementary level, the aim is more to the knowledge of the foreign language. In the case of level A, such errors do not influence so much the final result of a text produced by non-Italian users. On the contrary, for level B, it is necessary to use Italian words, if possible, less frequent vocabulary and words written in a grammatically correct way. Appropriate language means the use of words in the Italian language and not from a lexical point of view. In other words, we find certain written productions in which words are used that resemble other Italian words, but are from another language (for example *honora, phonetics*). These words are not part of the Italian vocabulary, but they are very reminiscent of Italian words. This part of the analysis belongs to the lexical sector, always based on the use of the Italian language.

Table 2. Total result of the variables "Appropriate language" and "Inappropriate language" according to the Language Level

Appropriate or inappropriate language * Language Level Crosstabulation				
		Appropriate or inappropriate language		Total
		Appropriate language	Inappropriate language	
Level	B1	98	62	160
	B2	88	68	156
Total		186	130	316

According to Table 2 and Fig. 1, among 316 texts produced there are some in which the Italian language is not used. On the contrary, we observe words with Greek characters or English, German, Greek, French, Spanish vocabulary. This phenomenon occurs perhaps because many Greek students unconsciously use interference by confusing certain words that are similar in different languages. According to these data, for the B1 level there is a total of 160 texts produced among which in 98 the Italian language is used and in 62, it is not used. With respect to the B2 level, there are 156 texts produced among which in 88 the appropriate language is visible and in 68 the inappropriate language is used.

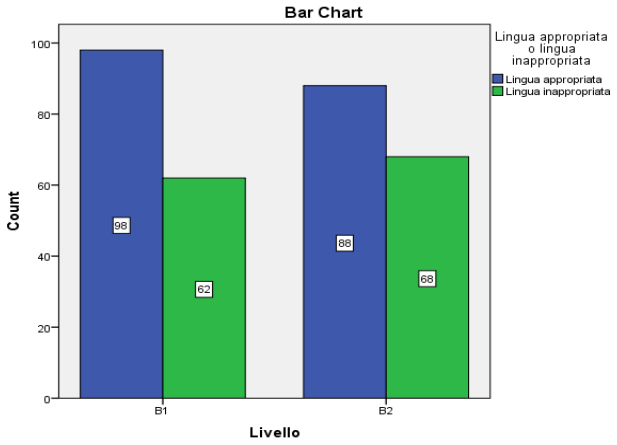

Fig. 1. Illustration of the total result of the variables "Appropriate language" (blue color) and "Inappropriate language" (green color) according to the Language Level (Livello)

Furthermore, in many texts we see words that exist in other languages that are similar to Italian words and others that do not exist in either the other language or in Italian. This is a frequent phenomenon because many Greek candidates try to combine and create words between two or more different languages. In other cases, we see Greek letters in Italian words that make the readability and comprehensibility of the text more difficult for a native Italian reader. In this way, the entire language level of the content worsens and the degree of difficulty is lowered. The same also happens when Greek names and surnames are used, for example, *Markos Papathanasiou* the name and last name are selected by case, it is only a virtual name and surname) or Greek cities and islands. Sometimes, many authors use several words in Greek instead of Italian. This contrast is noted for words such as Thessaloniki instead of Salonicco, Kerkyra instead of Corfù (for Italian language). Furthermore, in most texts we see a continuous repetition of words belonging to the VdB (Basic Vocabulary) through which the lexical density decreases which seems to lead to lower levels. According to the observations cited above, we see that both for the B1 and B2 levels there are many written productions that do not correspond to the relative level, since it will be difficult to evaluate a text that is not in Italian or without having the prerequisites for each language level. In some cases it is also difficult to understand the content when the sentences are very short or contain many foreign or non-existent words. A frequent situation is that many times we encounter words from another language in Italian. Such lexical elements can have the same meaning and sometimes another. These words, according to Russo, exist in English and Italian, they not only have the same meaning, but they are also similar in terms of their form as, for example, in the case of the words in the following table (Russo, 1998: 14).

Table 3. Examples of grammatical and lexical similarities between the English and the Italian language

English	Italian
Course	corso
University	università
Exam	esame

Consequently, if we think about this phenomenon as described in Table 3, it will be very easy for a non-Italian author to use words that have the same phonetics, but different semantics or are often confused due to interference. This happens perhaps because learning one or more languages in which the same vocabulary appears with a different spelling and different meaning, will be easier to confuse. The incorrect use of these words by Greek users leads to the inappropriateness of the language and, unfortunately, to the fact that they cannot be understood in Italian when they write a text in Italian. Furthermore, sometimes, words such as Greek names or surnames are used that do not exist at all in Italian. Even in this case there is a risk of being evaluated negatively, because this decreases the degree of readability of the text. That is to say, the measurement and evaluation of the texts produced is performed according to Italian rules. Consequently, it would be essential to emphasize that when writing one must reflect in Italian and not Greek, thus avoiding errors such as those reported. Another important phenomenon is that sometimes Greek words are used with Greek characters, perhaps because users do not know how to write them in Italian. Even in this case, a native Italian speaker evaluator might not be able to understand the text.

Speaking in more detail, in Table 4 we see some of the words in some of the texts produced by Greek users that are not part of the appropriate language with respect to the number of the written test and the Language level:

Table 4. Some very important errors found in texts produced in Italian by Greek users

B1 Language Level:
Incontrarà, 150.000 turistes, Messogia, grecasi presedanno, Visitore, vuò fare, Attivite, lebberi, Cantauotore, Politismo, offrè, attivitè, Parko, la città, que, cuando, attiviti del'arte, attiviti, Di Ampiente, piu du , laografico, exposizione,
B2 Language Level:
Phonetics, honora,risponsibile, endirizia, speridate, Supporte, organizato, attivite, Septembre, Visitore, Laografia, Messogeia, il curso, que conseguito, Specificarò, Mesogia, Fando, incontrerate, Jiugno, Visitori, vuò rilassare, piadare, Curso, senderlo, senderlo, Differento, oppurtunità, Qualre, physik, Visitore, greka, visitori, Idela, altrenative

A very significant factor is the lexical density contained in a produced text because the written language constitutes a permanent, static and durable sector. Moreover, the written language presents a higher recurrence from the lexical point of view. The lexical density appears indicative of a deeper process, if we talk about the relationship between lexical and functional words in a text or textual collections. It is linked to the vocabulary and the known words. The balanced lexical density, according to free mediations in English, is about 20% to 50% (Stegen, 2005: 8). This theory means that half of each sentence is composed of lexical words and functional lexical means. A text with low density will have less than 50% and a text with high density more than 50%. Academic and political texts tend to produce a higher density.

In Table 5, we see that there is no correlation between textual difficulty and lexical density because it marks -0.16%. The more difficult a text is, the higher the density and inversely proportional, that is, the lower the density the easier a text is.
Since lexical density plays an important role in general, textual difficulty depends on the vocabulary used in a total text. In this respect, if we use rarer and less known scientific words, the lexical density will be higher. Consequently, the level of textual difficulty will also be higher. There is always a reciprocal relationship between words and difficulty, as we see within these written productions. In fact, we often find in produced texts words that are repeated and belong more to the A1 and A2 level or Italianized words that are Greek, as happens with one's own names and surnames.

Table 5. Correlation between the Global Read-It Index (degree of text difficulty) and lexical density

Correlations			
		Read-It Global	Lexical density
Read-It Global	Pearson Correlation	1	-,016
	Sig. (2-tailed)		,778
	N	316	316
Lexical density	Pearson Correlation	-,016	1
	Sig. (2-tailed)	,778	
	N	316	316

In Table 6 and Fig. 2, the correlation between the Global Read-It difficulty level and the language used in the written productions is shown depending on whether the vocabulary is or is not in Italian language. In some cases, words that are not Italian are used, but which resemble the Italian lexicon. On the other hand, we observe in the majority of the texts produced (116 written productions) 1.01-7.60%. Thus, in 73 texts, the Italian language is used appropriately and in 43 texts, words from Greek are confused in Italian or even from other foreign languages, such as English or French (for example, the word *Athene* instead of Athens). This phenomenon is more observed in the variable "Orthography" (spelling), but it is certainly taken into account also in the present case of the inappropriate language. Furthermore, the more the Global Read-It percentage increases, the fewer texts produced we observe in which the appropriate language is followed. This result could be justified, because Greek writers find it difficult to memorize the right words in Italian. In fact, it is probably obvious not to find texts, in which the appropriate language is used 100%, since their producers are not of Italian origin.

Table 6. Total result of the variables "Appropriate language" and "Inappropriate language" according to the difficulty level Read-It Global

Read-It Globale (Binned) * Lingua appropriata o lingua inappropriata Crosstabulation				
Count				
		Appropriate or inappropriate language		Total
		Appropriate language	Inappropriate language	
Read-It Global (Binned)	<= 1,00	36	22	58
	1,01 - 7,60	73	43	116
	7,61 - 14,20	20	17	37
	14,21 - 20,80	9	11	20
	20,81 - 27,40	16	9	25
	27,41 - 34,00	9	9	18
	34,01 - 40,60	3	3	6
	40,61 - 47,20	3	3	6
	47,21 - 53,80	0	1	1
	53,81 - 60,40	4	5	9
	60,41 - 67,00	3	0	3
	67,01 - 73,60	4	1	5
	73,61 - 80,20	4	2	6
	86,81 - 93,40	1	2	3
	93,41 - 100,00	1	2	3
Total		186	130	316

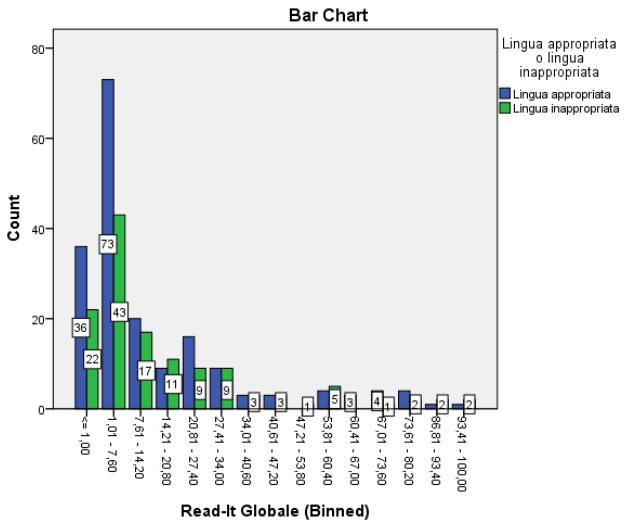

Fig. 2. Illustration of the final result of the variables "Appropriate language" (blue color) and "Inappropriate language" (green color) according to the Global Read-It difficulty level

Conclusion

After examining all these elements mentioned above, we can differentiate between the language levels in the texts produced by Greek users of the Italian language with respect to grammatical, lexical and syntactic factors. After the factor analysis, there is a high correlation between the components "Evaluation", "Global Read-It", "Spelling" and "Appropriate Language or Inappropriate Language". If we consider that the average should be more than 0.5% for Greek users, all these factors mark high values and are of great importance for both language levels, B1 and B2. Finally, having now new factors regarding texts produced by foreign candidates, it is expected from other colleagues and future researchers to discover similar or equal characteristics also for language levels A and C. Even more useful would be a broader research in the other parts of an entire exam (for example, in the oral part). A broad collection of data from additional units will give the possibility to the constructors of exams for different certifications to produce tests that will be fair and valid for foreign examinees, but without excluding that through further research. Such factors could also provide more information in producing fair and valid tests for Italian users of various foreign languages.

Statement of competing interests: The author has no competing interests.

REFERENCES

1. Chan, S. H. C., 2013. *Establishing the validity of Reading-into-Writing test tasks for the UK academic context.*Retrieved 6/2/2020, from shorturl.at/abtxL.
2. Dell'Orletta, F., Wieling, M., Cimino, M., Venturi, G., Montemagni, S., 2014. Assessing the Readability of Sentences: Which Corpora and Features? In *Association for Computational Linguistics*, 163-164.
3. Dell'Orletta, F., Montemagni, S., Venturi, G., 2011. READ–IT: Assessing Readability of Italian Texts with a View to Text Simplification. In *Association for Computational Linguistics*, 73-83.
4. Frigo, M., Zuppiroli, M., Pagani, R., 2007. *L' italiano L2 per le discipline: Lingua della socializzazione e Lingua dello studio.* Bologna: Centro Documentazione/Laboratorio per un' educazione interculturale. Retrieved 7/2/2020, from shorturl.at/eftBM.
5. Gunn, S. R., 1998.Support Vector Machines for Classification and Regression. In *ISIS Technical Report*, 1.
6. Johansson, V., 2009. *Developmental Aspects of Text Production in Writing and speech.*Lund: Lund University.
7. Lyding, V., Brunello, M., Dittmann, H., Stemle, E., Castagnoli, S., Lenci, A., Borghetti, C., Dell'Orletta, F., Pirrelli, V., 2014. The PAISA Corpus of Italian Web Texts.In *Association for Computational Linguistics*, 39.
8. Mariani, L., 2014. *Il Quadro Comune Europeo di Riferimento e la sua valenza formativa*. Retrieved27/3/2020, from shorturl.at/nqIR6.
9. Matthiae, C., 2010. Valutazione della produzione scritta: parametri, griglie e soggettività. In *Italiano LinguaDue*, 1, 105.
10. Moss, P. A., 1994. Can There Be Validity Without Reliability? In *Educational researcher*, 6.
11. Panizza, S., 2016. *Profili attuali di qualità degli atti normativi e amministrativi.* Pisa: Pisa University Press srl.
12. Russo, G. A., 1998. *A conceptual fluency framework for the teaching of Italian as a second language.* Retrieved 27/3/2020, from shorturl.at/qsEK8.
13. Spiro, R. J., Taylor, B. M., 1980. On Investigating Children's Transition from Narrative to Expository Discourse: The Multidimensional Nature of Psychological Text Classification. In *The National Institute of Education. Technical Report*, 195, 1-48.
14. Stegen, O., 2005. Editing Rangi narratives: A pilot study in literature production. In *Edinburgh Working Papers in Applied Linguistics*, p. 8.
15. Tong, S., Koller, D., 2001. Support Vector Machine Active Learning with Applications to Text Classification.In *Journal of Machine Learning Research*, 1-22.
16. Venturis, A., [Βεντούρης, Α.], 2020. Η διαμόρφωση κοινωνικής αντίληψης σχετικά με την αξία μιας γλώσσας, ως εργαλείο επιβολής γλωσσικής πολιτικής: Η περίπτωση της ιταλικής στην Ελλάδα. *Ο πολιτικός και παιδαγωγικός λόγος για την ξενόγλωσση εκπαίδευση.* Αθήνα: Πεδίο.
17. Venturis, A., (2018). *La selezione di testi italiani per il controllo della comprensione scritta: indicatori di livello linguistico e di difficoltà.* In Pirvu, E., (a cura di), (2017). Il tempo e lo spazio nella lingua e nella letteratura italiana. Firenze: Franco Cesati Editore.
18. Zawacki, T. M., Habib, A. S., 2014. *Negotiating "errors" in L2 writing: Faculty dispositions and language difference.* Retrieved 11/2/2020, from shorturl.at/xyG57.

****