





Developing LEMI: A new corpus based literacy support tool for schoolchildren

Roxana Rogobete^a, Alexandru Oravițan^b, Mădălina Chitez^c and Karla Csürös^d

^aDepartment of Romanian Studies, West University of Timisoara, , roxana.rogobete@e-uvf.ro; ^bDepartment of Modern Languages and Literatures, West University of Timisoara, , alexandru.oravitan@e-uvf.ro; ^cDepartment of Modern Languages and Literatures, West University of Timisoara, , madalina.chitez@e-uvf.ro and ^dDepartment of Modern Languages and Literatures, West University of Timisoara, , karla.csuros@e-uvf.ro

How to cite: Rogobete, R.; Oravițan, A.; Chitez, M.; Csürös, K. (2023). Developing LEMI: A new corpus based literacy support tool for schoolchildren. In *CALL for all Languages - EUROCALL 2023 Short Papers*. 15-18 August 2023, University of Iceland, Reykjavik. <https://doi.org/10.4995/EuroCALL2023.2023.16966>

Abstract

This study centres on the developing and testing stages of a literacy support tool dedicated to young schoolchildren. The LEMI tool is currently in development (since January 2023) at the CODHUS research centre (Centre for Corpus Related Digital Approaches to Humanities) from the West University of Timișoara, Romania. LEMI aims to stimulate interest in reading during the first individual and collective reading activities (ages 7-11). This aim will be achieved by creating a digital reading repository with a user-friendly interface that verifies reading text complexity and delivers automatic reading-level reports to users. We use corpus linguistics methods to create a text complexity formula adapted to the Romanian language, which can be integrated into the automated complexity evaluation interface in LEMI. The necessity of such an instrument is motivated by the fact that, in Romania, there are increased rates of functional illiteracy and school dropout. We hypothesise that texts must be level-adapted (according to grade or readability) for schoolchildren to relate positively to reading activities. In the Romanian context, LEMI is the first digital tool wholly tailored to children's literature, which complements national curricula and didactic materials provided to young children. Distinctively, LEMI responds to the need for easy-to-use tools to adapt reading individually, according to the reader's profile. LEMI is a unique tool, not only for Romanian but also for children's literature in other languages. The functionalities of the LEMI pilot version will be tested with the partners involved in the project (three schools from Timiș county and an educational NGO).

Keywords: *LEMI, literacy support tool, corpus based literacy tool, children's literature repository, text complexity automatic assessment, text complexity in Romania, readability for Romanian language.*

1. Introduction

Understanding the process of reading comprehension and developing reading proficiency is essential from a variety of perspectives, from the cognitive to the pedagogical and cultural approach. According to the OECD (2009, 2), reading literacy is defined as “understanding, using, reflecting on, and engaging with written texts, in order to achieve one’s goals, to develop one’s knowledge and potential, and to participate in society.” Students’ engagement is a key aspect of lifelong learning (Ho & Lau 2018), yet motivating children to pursue reading has proven to be a challenge in recent decades, as the digital transformation has become a commonplace.

The last PISA reports (n.d.) focused on 15-year-old students, and the assessment of their key knowledge and skills show an almost negligible decrease in the students' reading skills from 2015 to 2018 (all countries taken into consideration, see Figure 1).

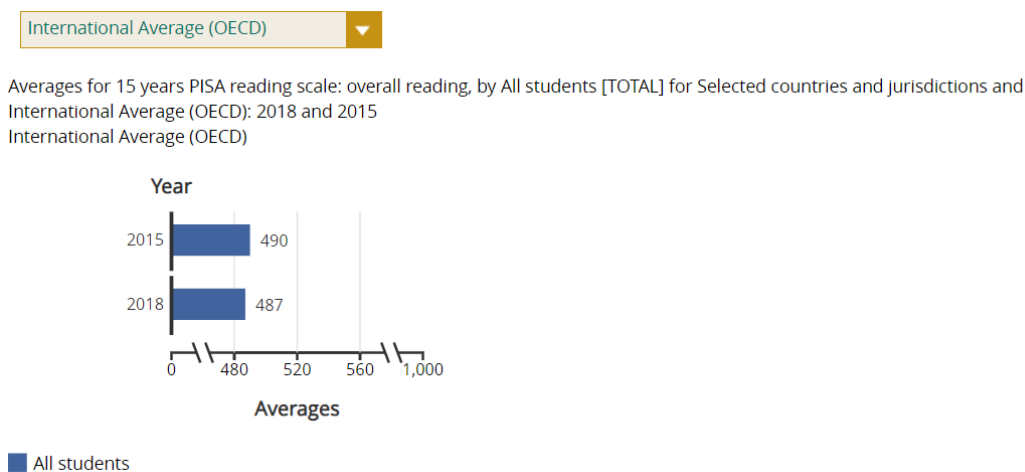


Figure 1. PISA reading scale: overall reading (All students, all countries).

<https://pisadataexplorer.oecd.org/ide/idepisa/report.aspx>

However, when the same criteria are selected for a country such as Romania alone, the scores are lower in all the subjects discussed (reading, mathematics, science, see OECD 2019, 1). Not only are the numbers lower in the case of Romanian students (with approximately 11-12%), but the decline is also faster (Figure 2): the reading attainment scores are overall lower for Romania and the difference between 2018 and 2015 is greater than the average for all students (as seen in Figure 1).

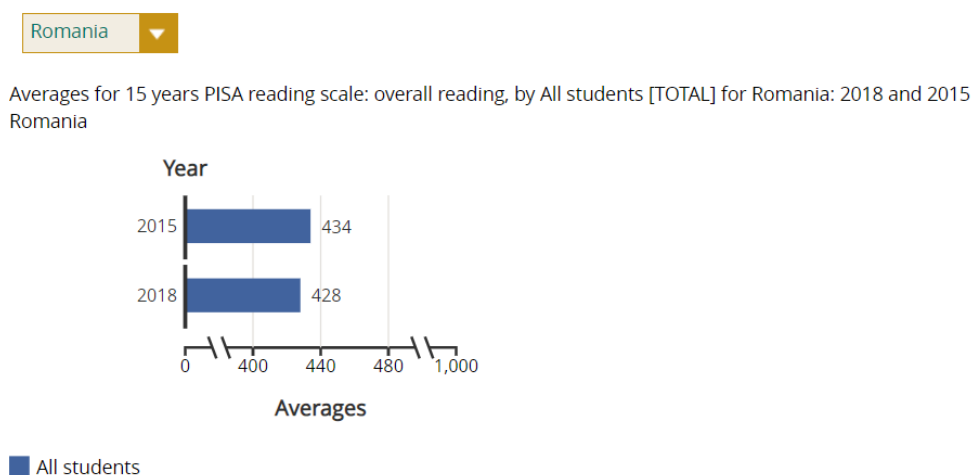


Figure 2. PISA reading scale: overall reading (All students, Romania).

<https://pisadataexplorer.oecd.org/ide/idepisa/report.aspx>

Over the last decade, Romania has seen an increase in functional illiteracy and school drop-out rates; the country has the highest rate of early school leavers in Europe (15%, compared to less than 10% at the European level, see Eurostat 2022, 1; and European Commission, 2022, 3), despite the fact that numerous initiatives have been taken to improve reading behaviour in both the 'traditional' and digital environments. Thus, the school curriculum assumes a direct link between reading activities and increasing the motivation to read. However, the lack of a rigorous way of structuring information based on the criterion of cognitive progression and systematicity

produces the opposite effect: decreased motivation to read. Despite the wide range of school textbooks approved by the National Center for Policies and Evaluation in Education (NCPEE), there are no unified school reading recommendations based on linguistic research, which have measured the complexity of the recommended texts (which in many cases have remained unchanged for decades) by reference to reading level (readability, in the technical literature) specific to each age category. Readability is defined as the degree to which a given group of people find a certain piece of reading comprehensible (McLaughlin, 1969) or as the ease with which a text can be understood due to its style of writing (Klare, 1963). In this context, there is a need to develop digital resources and tools tailored to the specific demands of distinct group ages or school levels. Although educational curricula and selected texts present in textbooks have to be related with research regarding measurement of text complexity, few children’s reading apps have been developed, and the majority is limited to English language texts (see Chitez et al., 2023). In Romania, there are no school reading series whose content has been assessed at the level of text complexity, but there is one reading application that contains digital versions of literary texts. However, they are selected using human recommenders, without any linguistic analysis or automatic assessment.

2. LEMI – a corpus based literacy support tool

While readability studies in English are not new, research for other languages such as Romanian is scarce. However, studies such as Botarleanu et al. 2023 have taken the first steps towards developing methods in order to measure the complexity of words within texts across different languages (p. 2). This is where the LEMI project (*Reading for Me. Science for Children*, coordinated by the West University of Timisoara, Romania) will fill in the research gap and become “the first Romanian tool that uses computational linguistics methods to assess school children’s literature complexity and readability” (Chitez et al., 2023, 2), in order to achieve a “correct pairing of textbook complexity and student grade level” (Paraschiv et al. 2023, 52). LEMI’s technical profile is a SaaS that uses Machine Learning (ML) and Natural Language Processing (NLP) methods to automatically assess young children (age 7-11) literature’s complexity, readability, and age adaptability. Moreover, it will offer an online platform with a digital repository of school reading, useful for children, teachers, NGOs, and other stakeholders, who will have immediate access to any new forms of written culture. Not only will the available literary corpus be linguistically assessed, but LEMI will provide an automated evaluation of the reading level for texts selected/uploaded by users.

3. Testing readability formulas for Romanian

The main challenge in this context is to develop the criteria for the automated assessment. In order to test the available readability formulas, the research team selected texts already included in school curriculum (several texts present in textbooks designed for different classes). For instance, an excerpt from a classical author such as Ion Creangă (adapted from *Amintiri din copilărie*) was included in a 3rd grade textbook (Figure 3).

Amintiri din copilărie
– fragment –
după Ion Creangă

Vocabular

iarmaroc: târg;
a o tuli: a fugi;
tolănit: întins într-o poziție confortabilă;
prund: mal, țăr, teren acoperit cu pietriș;
tupilus: pe furis;
păpușoi: loc plantat cu porumb.

Într-o zi, pe aproape de Sânt-Ilie, se îngămădise, ca mai totdeauna, o mulțime de treburile pe capul mamei. Și mă scoală mama atunci, mai dimineață decât alte dăți, și-mi zice cu toată inima:
— Nică, dragul mamei! vezi că tată-tău e dus la coasă, căci se scutură ovăzul, și eu nu-mi văd capul de treburile! Tu mai lasă drumurile și stai lângă mămuca, de leagănă copilul; c-apoi și eu ți-oi lua de la **iarmaroc** o pălăriuță cu pană și o curălușă!
— Bine, mamă! Dar în gândul meu numai eu știam.
Când auzeam de legănat copilul, nu știu cum îmi venea; căci tocmai pe mine căzuse păcatul, să fiu mai mare între frați. Însă ce era să faci, când te roagă mama?
Dar în ziua aceea, în care mă rugase ea, era un senin pe cer și așa de frumos și de cald afară, că-ți venea să te scalzi pe uscat, ca găinile. Văzând eu o vreme ca asta, am **tulit**-o la baltă.

Figure 3. Textbook excerpt, 3rd grade, Aramis publishing house.

<https://manuale.edu.ro/manuale>

Readability formulas that are widely used for English and available on free online platforms delivered inconclusive results, because the same text is classified in multiple ‘grades’ or ‘classes’ (see Table 1). All formulas focus on linguistic factors, such as word length and sentence length: The Flesch Reading Ease formula (Flesch, 1948) and the Gunning Fog Index (Gunning, 1952) use sentence length; the Simple Measure of Gobbledygook (McLaughlin, 1969) takes into account the number of syllables per word and number of polysyllabic words; while the Coleman-Liau Index (Coleman & Liau, 1975) and the Automated Readability Index (Smith & Senter, 1967) quantify characters instead of syllables per word. Using the metrics available online not only confuses the reader in terms of text classification, but the calculations are not appropriate for Romanian, as the manual scores and calculations show in Table 1.

Table 1. Online available automated readability calculations.

Formula	Manual score	Online score provided by readabilityformulas.com
Flesch Reading Ease (FRE)	50.81244	72.8
Gunning Fog Index (GFI)	15.75385	9.7
Coleman-Liau Index (CLI)	7.503671	1
Simple Measure of Gobbledygook (SMOG)	8.196152	6.5
Automated Readability Index (ARI)	7.705217	2.5

When manually calculating the number of words, syllables, letters, and sentences in the excerpt (see Table 2), we concluded that the main problem with the available readability formulas is quantifying the number of syllables in a low-resource language like Romanian. Hence, the main challenge in creating the support tool for Romanian will be to integrate an appropriate syllable separator and counter into an existing readability formula.

Table 2. Manual calculations of syllables and letters.

Sentence (separator . ;)	Words	Syllables	Letters (without hyphens)
1	18	35	82
2	15	29	65
3	3	6	15
4	18	27	68
5	11	22	51
6	15	26	55

7	2	4	8
8	7	10	26
9	10	17	43
10	12	20	51
11	9	13	30
12	31	48	107
13	10	16	36
TOTAL	161	273	637
AVERAGE	12.38461538	21	49
Words with 3 or more syllables	27		
syllables/words	1.695652174		
letters/words	3.956521739		

4. Discussion and conclusions

The lack of consistency provided by existing readability formulae requires researchers to develop specific tools for the Romanian language. Which of the existing formulae is the most appropriate? Are any of them suitable for Romanian? What would a specific formula for Romanian look like? How do we define the reporting scale/scheme? – these are key questions for designing LEMI’s main automated formulae. Since the currently available options are not sufficient, the development of a new readability formula for Romanian will take into account ‘traditional’ criteria, such as text, sentence and word length, as well as alternative, but necessary measures, such as “word maturity”, “age of exposure word lists” (see Botarleanu 2023). The latter aspects will focus more on comprehension and text complexity in terms of semantics, while the traditional metrics will address the morphological and, more generally, the grammatical dimension of a text. Moreover, classifying children’s literature with the help of BERT-based models (Bidirectional Encoder Representation from Transformers) (Paraschiv et al. 2023) will enhance the innovative character of LEMI, by providing the testing and scientific validation of the first method of assessing the complexity of school texts in Romanian. The computational validation of a new formula within the ReaderBench framework will be complemented by a process of ‘user validation’, where students will have access to a beta version of the LEMI platform and will be able to individually rate the texts from the compiled corpus. This hybrid, multi-stage validation will refine the development of an exclusive concept of technological transfer from the scientific area to the educational one, in the field of early school reading.

Acknowledgements

PROIECT CO-FINANȚAT DE:



This work was supported by a grant of the Administration of the National Cultural Fund (AFCN) of the Romanian Ministry of Culture, in the framework of the programme *Promotion of Written Culture*, session I/2023, for the project LEMI (*Lectură pentru mine. Știința în slujba copiilor – LEMI/ Reading for Me. Science for Children*; January-November 2023), contract no. P0299/10.02.2023. The project was awarded to the West University of Timisoara for the proposal submitted by the project coordinator, CS II Dr. Habil. Madalina Chitez.

References

- Botarleanu, R. M., Dascalu, M., Watanabe, M., Crossley, S. A., & McNamara, D. S. (2022). Age of Exposure 2.0: Estimating word complexity using iterative models of word embeddings. *Behavior research methods*, 54(6), 3015–3042. <https://doi.org/10.3758/s13428-022-01797-5>.
- Chitez, M., Rogobete, R., & Oravitan, A. (2023, June). Designing LEMI: the Romanian language tool that makes kids love reading. In *Conference Proceedings. The Future of Education 2023*. <https://conference.pixel-online.net/files/foe/ed0013/FP/3177-PRI6037-FP-FOE13.pdf>.
- Coleman, M. and Liao, T. L. (1975). A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2), 283–284. <https://doi.org/10.1037/h0076540>.
- DuBay, W. H. (2004). *The principles of readability*. Online Submission. Available at: <https://files.eric.ed.gov/fulltext/ED490073.pdf>.
- European Commission, Directorate-General for Education, Youth, Sport and Culture (2022). *Education and training monitor 2022: Romania, Publications Office of the European Union*. Retrieved October 10, 2023, from <https://data.europa.eu/doi/10.2766/310121>.
- Eurostat (2022). *Early leavers from education and training. Statistics Explained*. Retrieved October 10, 2023, from <https://ec.europa.eu/eurostat/statistics-explained/SEPDF/cache/1150.pdf>.
- Flesch, R. (1948). A new readability yardstick, *Journal of Applied Psychology*, 32(3), 221–233. <https://doi.org/10.1037/h0057532>.
- Gunning, R. (1952). *The technique of clear writing*. McGraw-Hill.
- Ho, E. S. C., & Lau, K. (2018). Reading engagement and reading literacy performance: effective policy and practices at home and in school. *Journal of Research in Reading*, 41(4), 657–679. <https://doi.org/10.1111/1467-9817.12246>.
- Klare, G. R. (1963). *The measurement of readability*. Ames, Iowa: Iowa State University Press.
- McLaughlin, G. H. (1969). SMOG grading: A new readability formula. *Journal of Reading*, 12(8), 639–646.
- OECD (2009). *PISA 2009 reading framework*. Paris: OECD Publications.
- Paraschiv, A., Dascalu, M., & Solnyshkina M. I. (2023). Classification of Russian Textbooks by Grade Level and Topic Using Readerbench. *Научный результат. Вопросы теоретической и прикладной лингвистики (Research Result. Theoretical and Applied Linguistics)*, 9(1), 73–86. <https://doi.org/10.18413/2313-8912-2023-9-1-0-4>.
- PISA IDE. (n.d.) *Averages for age 15 years PISA reading scale: overall reading, by All students [TOTAL] and jurisdiction: 2018 and 2015*. Retrieved July 31, 2023, from https://pisadataexplorer.oecd.org/ide/idepisa/report.aspx?p=1-RMS-1-20183.20153-PVREAD-TOTAL-IN2.IN3-MN_MN-Y_J-0-0-37&Lang=1033.
- Smith, E. A., & Senter, R. J. (1967). *Automated readability index*. Aerospace Medical Research Laboratories, Aerospace Medical Division, Air Force Systems Command. Wright-Patterson Air Force Base, Ohio.
- Ziming, L. (2005) Reading behavior in the digital environment. *Journal of Documentation*, 61(6), 700–712. <http://dx.doi.org/10.1108/00220410510632040>.