Vocabulary diversity of reading materials at the beginning of primary school

Katarina Aladrović Slovaček
University of Zagreb, Croacia

https://doi.org/10.36505/TheLinguisticProceedings/2025/17/02/020/000706

Abstract

Vocabulary represents the most dynamic aspect of language development, continuously changing, expanding, and being acquired throughout life. Lexical diversity serves as a predictor of academic achievement and is a hallmark of competent communication in adulthood. Increasingly, there is concern regarding the limited understanding of word meanings among younger school-aged children, while less attention is given to how vocabulary development and lexical diversity can be systematically monitored. This paper therefore analyses all mandatory reading materials for first- and second-grade primary school students to calculate lexical density and diversity, defined as the ratio of the number of variants to the total number of occurrences. It is hypothesized that unfamiliar and archaic words appear frequently, and that sentence structures may be more complex than expected for this age group. The data gathered yield a frequency dictionary that may be applied in constructing tests for assessing lexical diversity and competence in early school-age children.

Keywords: obligatory reading books, textbooks, Croatian language, reading, lexicon

Introduction

Vocabulary development plays a crucial role in overall language acquisition, as it unfolds throughout an individual's entire life. In early childhood, preschool, and the initial school years, vocabulary growth serves as a key indicator of linguistic development and a predictor of educational success and later academic skills (Radić et al., 2010). In Croatian, as a first language, there are no standardized tools for measuring vocabulary development; rather, it is typically estimated based on researchers' observations. It is generally assumed that a child's mental lexicon contains approximately one thousand words at the age of three, four thousand by the age of four, and around ten thousand by the age of six or seven—when entering school (Pavličević-Franić, 2011). Naturally, learning words in one's first language differs substantially from learning words in a second language, primarily because exposure in the first language occurs through both spoken and written modalities (Webb & Nation, 2017). Similarly, Biber and Conrad (2001) emphasize that lexical competence relies on pragmatic, syntactic, morphological, and phonological knowledge, particularly in the case of the first language.

The impact of sociocultural factors is confirmed by Southwood et al. (2021), who identified gender, age, environment, and initial education as key predictors

of lexical diversity. Moreover, Duff et al. (2015) demonstrated that reading habits and experiences not only expand vocabulary size but are also closely linked to literacy development. Song et al. (2015) further notes that individual differences in vocabulary breadth and richness are shaped by cognitive skills—especially working memory and morphological awareness—as well as by the home environment. Ultimately, as Webb and Nation (2017) stress, words form the very structure of language and hold a central position in all linguistic activities—listening, speaking, reading, and writing—while also constituting an essential part of everyday communication. In traditional education, vocabulary teaching was often neglected (Duan & Da, 2015), with greater emphasis placed on grammar and orthography than on lexical acquisition. Yet, even in first-language contexts, vocabulary learning is of paramount importance.

Books read in childhood profoundly shape vocabulary acquisition, and reading habits substantially influence vocabulary growth, as shown by Cvikić (2007), who found that school-age children acquire between 500 and 1,000 new words each year. To fully acquire a new word, children typically need to encounter it seven to ten times in different contexts (Laufer, 2005), while being encouraged to use it in both spoken and written forms. Reading, as a secondary linguistic activity, plays a particularly important role by exposing children to new information and enabling implicit vocabulary learning through text.

Research methodology

This study examined books designated as mandatory reading according to the *National Curriculum for the Croatian Language* (2019), intended for students at the beginning of schooling—specifically, in the first and second grades of primary education within the subject *Croatian Language*, which serves as students' first and native language. The analyzed works included fairy tales by the Brothers Grimm (*Little Red Riding Hood*, *Snow White*, and *Sleeping Beauty*) and those by Hans Christian Andersen (*The Emperor's New Clothes*, *The Ugly Duckling*, and *The Daisy*). In addition to these literary works, the study also included textbooks from three publishing houses used in the first and second grades, covering *Croatian Language*, *Science and Society*, and *Mathematics*. The texts were processed using **Sketch Engine**, a software tool for computational text analysis. The quantitative data obtained were subsequently analyzed and compared using **SPSS**. The research questions were: **P1 and P2:** To examine lexical diversity and density in the corpus of required reading and textbooks for the first and second grades of primary school. **P3:** To examine the number of words per sentence in the corpuses; **P4:** To examine the ratio of low-frequency to high-frequency words in the corpuses.

The first research goal was to determine lexical density and diversity within the mandatory reading materials assigned to students at the beginning of schooling. The results revealed that these texts contain approximately **1,020 lemmas**, **1,621 distinct word forms**, and **4,014 tokens**. Lexical diversity, expressed

as the ratio of distinct word forms to the total number of words, was **0.40**, while lexical density, defined as the ratio of lemmas to tokens, was **0.25**. These findings indicate that the texts are lexically rich, with approximately **25%** of words recurring, while **48%**of words occur only once.

The second objective of the study was to examine lexical density and diversity in textbooks for *Croatian Language*, *Mathematics*, and *Science and Society* published by three different publishing houses. Out of a total of **51,125 words**, the number of distinct word forms was **12,354**, and the number of lemmas was **6,335**. Lexical diversity across textbooks for all three subjects amounted to **0.12**, while lexical density was **0.24**. Approximately **35%** of words were repeated, and **52%** appeared only once. These findings suggest that textbooks are lexically less demanding than mandatory literary works, as the number of recurring words is significantly higher in textbooks compared to required readings. When comparing the textbook corpus with the corpus of mandatory reading materials, it becomes evident—as expected—that a greater variety of words occurs in the literary texts. As previously discussed, reading plays a crucial role in vocabulary development; therefore, a larger number of unfamiliar words and a richer lexical structure in literary works naturally contribute to vocabulary expansion among students.

The third objective of the study was to analyze the sentence structure of the texts. In the textbook materials, a total of **5,232 sentences** were identified, while the corpus of literary works contained **169 sentences**. By dividing the total number of tokens by the number of sentences, it was determined that the average sentence length in the textbook corpus was **9.77 words per sentence**, whereas in the corpus of literary works it was **9.59 words per sentence**. These results indicate that sentence length in both corpora is comparable and appropriate for children aged seven to nine years, corresponding to the expected developmental level at the beginning of schooling.

The fourth research objective was to identify the words that occur most frequently in the required reading texts and in the textbooks. As expected, **nouns** dominate in the literary corpus, often serving as the key elements of the stories—for instance, *queen*, *dwarf*, *mirror*, *Snow White*, *seven*, and the adjective *beautiful*. In contrast, **verbs** predominate in textbook materials, typically serving instructional functions such as directing or engaging the learner—for example, *to know*, *to develop*, *to write*, and similar forms.

Conclusion

The analysis confirmed that vocabulary plays a crucial role in understanding both literary and textbook materials. At the beginning of schooling, it is particularly important to select words carefully for the texts used in instruction. Equal attention should also be devoted to the vocabulary of mandatory reading materials, as these texts have a significant influence on the process of vocabulary enrichment during early education.

The findings of this study revealed that some texts are excessively complex for students' comprehension levels, highlighting the necessity of monitoring and adjusting lexical complexity. High lexical difficulty can substantially hinder text comprehension and, consequently, students' overall learning success.

Therefore, the vocabulary used in both textbooks and literary works should be appropriate to students' developmental stages, enabling children to expand their lexicons through exposure to suitable linguistic input. Such materials not only facilitate knowledge acquisition but also contribute to the development of linguistic confidence and communicative competence.

References

Biber, D., Conrad, S. 2001. Quantitative corpus-based research: Much more than bean counting. TESOL Quarterly, 35, 331-336.

Cvikić, L. and com. 2007. Drugi jezik hrvatski. Profil: Zagreb.

Duff D., Tomblin J.B., Catts H. 2015. The Influence of Reading on Vocabulary Growth: A Case for a Matthew Effect. J Speech Lang Hear Res. 58(3): 853-64. doi: 10.1044/2015_JSLHR-L-13-0310.

Duan, J., Da, H. 2015. Semantics and Vocabulary Acquisition and Teaching. U: Studies in Literature and Language,10 (6), 67-71.

Laufer, B. 2005. Focus on Form in second language vocabulary learning, EUROSLA Yearbook 5, 223-250, John Benjamins Publishing Company: Amsterdam: Philadelphia.

Kurikulum nastavnoga predmeta Hrvatski jezik. 2019. Zagreb: MZO.

Pavličević-Franić, D. 2011. Jezikopisnice. Alfa: Zagreb.

Radić, Ž. and com. 2010. Udžbenik kao poticaj ili prepreka leksičkomu razvoju. Lahor, 9, 43-59.

Song S., Su M., Kang C. Liu H., Zhang Y., McBride-Chang C., Tardif T., Li H., Liang W., Zhang Z., Shu H. 2015. Tracing children's vocabulary development from preschool through the school-age years: an 8-year longitudinal study. Dev Sci. Jan;18(1):119-131. doi: 10.1111/desc.12190. Epub 2014 Jun 24. PMID: 24962559; PMCID: PMC4276547.

Southwood F., White M.J., Brookes H., Pascoe M., Ndhambi M., Yalala S., Mahura O., Mössmer M., Oosthuizen H., Brink N., Alcock K. (2021). Sociocultural Factors Affecting Vocabulary Development in Young South African Children. Front Psychol. doi: 10.3389/fpsyg.2021.642315.

Webb, S., Nation, P. 2017. How Vocabulary is Learned. Oford University Press: Oxford.