Croatian Journal of Education
Vol.25; No.1, pages: 247-283
Preliminary communication
Paper submitted: 7th September 2021
Paper accepted: 16th September 2022
https://doi.org/10.15516/cje.v25i1.4569

Textbook Materials for the Lower Grades of Elementary School: Do They Conform to the Principles of *Plain Language*?

Ana Matić Škorić, Jelena Kuvač Kraljević and Mirjana Lenček
University of Zagreb, Faculty of Education and Rehabilitation Sciences

Abstract

Plain language is a concept that spread across Europe half a century ago. It refers to communication based on utterances and structure from which listeners and readers can easily extract and understand necessary information. This study presents the concept of plain language in the context of early education through the analysis of written materials to which pupils attending lower-grade elementary school are exposed. The aim of this study is to analyse phonological, lexical and syntactic characteristics of textbooks for first, second, third, and fourth grade and to link the obtained data with general principles of plain language. For the purpose of this study, a database of written texts was developed – School Corpus of Written Language (Riddys) – with 502 713 tokens. The distribution of graphemes, types of speech, as well as the lengths of words and sentences were analysed. The results show that of the three analysed language components, only the lexicon does not follow the principles of plain language. Too many words that occur only once in all written materials in the first two grades of elementary school do not contribute to faster and more efficient mastering of reading skills.

Keywords: *beginning reading and writing; elementary school education; linguistic analysis of textbooks; plain language; School corpus of written language (Riddys);*

Introduction

Plain language

Sixty years ago, the slogan "Language for All" spread across Europe to point out the difficulties that an average speaker of a language encounters every day in understanding different types of texts and messages (Maaß, 2020).

In response to the identified difficulties, *plain language* was introduced in the 1970s as a concept and tool that aimed not only to resolve them but also to raise awareness of the importance of reflecting on the appropriate means of language use (Arias-Badia & Matamala, 2020). As a counterpart to *plain language*, *plain communication* is also used because, according to authors such as Montolío (2019), it encompasses the scope of human interaction more broadly.

According to the International Plain Language Federation (IPLF), *communication is in plain language if its wording, structure, and design are so clear that the intended readers can easily find what they need, understand what they find, and use that information* (IPLF, 2021). The main objective of *plain language* is to make the content understandable to individuals it is intended for. Therefore, the slogan *less is more* illustrates its essence in the best possible way. *Plain language* is not only used for written language but also spoken language promoted in public, political, legal, and educational spaces, as well as the media (Arias-Badia & Matamala, 2020).

The most common cause for the inadequate understanding of a text or message is that language of a text is not adapted to the language knowledge of a reader or listener. For the text to be comprehensible, it must integrate several principles that ensure this, such as analyticity, directness and concreteness in writing, accuracy and systematicity in expressing the content, and emphasis on what is important (Maaß, 2020). This implies the use of frequent words with a clear and unambiguous meaning. The choice of syntactic structures and their length should be such that they can be retained in ‘ ‘one's short-term memory, and processed concisely and rapidly. The shaping of the text should be in line with the ‘ ‘person's cognitive level, and the topic should stimulate further interest and ensure a deeper and more critical reflection on it.

It is considered that the most common reasons for inappropriate use of language are the author's lack of understanding of the cognitive, and especially language abilities and needs of readers and users. Therefore, to shape a message that will successfully reach the user and be equally understandable to everyone, it is necessary to think constantly about what the reader can do and what they know. In this study, the concept of *plain language* will be presented in the context of early education, and it will be problematised through the analysis of written materials to which students attending lower grades of elementary school are exposed.[1]

Language in textbooks

Children who are just starting school significantly differ regarding the time and the manner of previous exposure to written language due to numerous internal and external factors such as family structure, family literacy, ‘ ‘child's enrolment in kindergarten

[1] The concept of plain language, apart from language structure of written and spoken communication, defines also the principles of graphic presentation and visual aspects of document design (see more in Arias-Badia and Matamala, 2000). However, these components of written materials are not relevant for this study and therefore will not be analysed.

or other formalised preschool programmes, as well as cognitive and psychological characteristics of a child (Celik, 2020). Enrolment into first grade ensures continual and daily exposure to written materials for all children and a stable environment for further language development. However, the speed and success with which children will understand the first written words are again determined by their cognitive and psychological characteristics, early reading and writing skills, teaching methods, and the characteristics of written materials on which formal instruction is based.

Even though the Ordinance on Textbook Standards (2013; 2019) clearly states that texts approved and certified for use by the Ministry of Science and Education (MSE) of the Republic of Croatia are appropriate for children regarding their cognitive and language development, and therefore encourage creativity and stimulate the development of critical thinking and metacognition, a series of studies shows the opposite. Unsatisfactory content of the first textbooks in relation to children's language abilities and literacy skills has been the focus of attention since the beginning of the 2000s when the first analyses of primers and characteristics of language to which children are exposed occurred (for example, Cvikić, 2002; Kovačević & Kuvač, 2004; Lenček & Gligora, 2010; Radić et al., 2010), as well comprehensive analyses of particular textbooks (e.g., textbook of Croatian Language; see Nemeth-Jajić, 2007). These studies mainly emphasise the inappropriate choice of words and syntactic structures that surpass a child's semantic knowledge and the possibility of language processing.

Three preconditions are necessary for stimulating vocabulary development through reading: well-developed grapheme decoding skills (a grapheme-phoneme connection); the ability to recognise unknown words; the ability to draw conclusions from the written context (Beck, McKeown & Kucan, 2002). Inadequate success in any of these preconditions will have a negative impact on further vocabulary and overall language development and will lead to a decrease in motivation for further exposure to written text (McGeown et al., 2016). Namely, there is a two-way link between reading, i.e. mastering this skill, and a motivation for reading, and this should not be neglected in the education process (Morgan & Fuchs, 2007). To maintain motivation and properly develop the skills, the content to which children are exposed should reflect their language development. According to Scott (2005), vocabulary development can be stimulated if children, beginner readers, are frequently exposed to familiar words, equally in repetitive and new contexts. Beck, McKeown and Kucan (2002) claim that one should be exposed to an unfamiliar word at least 12 times in its written form to understand it and acquire its meaning. An inappropriate ratio of unfamiliar and familiar words will have an unfavourable influence on vocabulary enrichment and, even more problematic, on the development of reading skills. Written language should be syntactically clear to children, and words should be suitable to their age and applicable in different contexts. The existing studies in Croatian have shown that this is not the case.

Kovačević and Kuvač (2004) found that more than 50 % of all words in primers occur only once. Furthermore, data on the level of similarity of different storybooks,

i.e. the data on the amount of linguistic content they share, reveal only a 30 % overlap in lexical context, with greater overlaps in the number of functional words than in the number of content words (e.g., nouns and verbs) (Cvikić, 2002). Studies that followed (Lenček & Gligora, 2010; Radić et al., 2010) have resulted in data almost identical to those from the beginning of 2000. One of the most recent studies on lexical content in primers (Miličić et al., 2017) included three randomly selected primers.[2] That study has once again confirmed that primers are too demanding and inappropriate, both regarding the occurrence of words (new and unknown), and regarding the ratio of low-frequency words. Materials to which children are exposed are neither in line with educational standards nor with psycholinguistic findings on the characteristics of language and literacy development.

During school age, under the influence of language development and exposure to written text, apart from vocabulary knowledge, syntactic forms (their length and type) also grow in complexity and diversity. While the Croatian literature reveals some data on the representation of sentence types in children's production (Radić Tatar, 2013) or their representation in primers (Miličić, 2016), there is a lack of comprehensive data on the length of such structures. Kovačević and Kuvač (2004) warn about the presence of extremely long sentences (with more than 40 words) in the first children's textbooks. This, together with the overabundance of words unfamiliar to children, contributes to the lack of understanding of the text. While analysing a spoken corpus, Vuletić (1991) showed that spoken sentences, on average, consist of 6.45 words, whereby the most frequently used sentences are those with four or five words.

The aforementioned data on lexical and syntactic structure of written materials arise exclusively from the analysis of materials used in the first grade of elementary school that have been a part of teaching materials during the last 20 years. The purpose of this study is to examine and compare written materials used in the lower grades of elementary school (first through fourth grades), on which the development of early and automatised reading and writing is based, and to determine whether nowadays they are more in line with the principles of *plain language*, whose planning starts from the understanding of cognitive and language abilities and needs of the reader.

Aim of the study

The aim of this study is to analyse phonological, lexical and syntactic characteristics of textbooks for lower elementary school grades and to compare the distribution of particular elements within the three mentioned language components in the first, the second, the third and the fourth grade. Furthermore, the aim is to link the obtained data on the characteristics of written materials with the general principles of *plain language*, as well as to discuss the extent to which the analysed textbooks reflect those

[2] In each school year the number of approved primers is different; in some school years the number is higher than seven.

principles, i.e. by having a satisfactory word frequency and length of syntactic structures. It is assumed that written materials attended for lower grades of elementary school will grow in complexity as the level of education increases, on the chosen measures: length of words measured by the number of graphemes, lexical diversity and length of sentences.

Methods

Analysed materials: School corpus of written language (Riddys)

For this study, a database of written texts from textbooks for elementary school students from four subjects was developed: Croatian Language, Mathematics, Science and Religious Education. In addition to the written materials in the subjects mentioned above, it also contains approximately 20 % of the required reading materials (i.e., book report titles), as well as several volumes of the educational periodicals for children approved by the MSE and procured by the school. All materials were transcribed and unified into the *School Corpus of Written Language* (Riddys; Kuvač Kraljević & Lenček, 2020; created specifically for the goals of the Riddys project). The list of textbooks, authors and publishers is provided in the Appendix.

The texts consist of 502 713 tokens (and about 45 000 sentences), and are divided into four subcorpora, one for each grade level. All four subcorpora were lemmatised and morpho-syntactically and syntactically tagged using the ReLDIanno tool, which is publicly available as part of the Slovenian project CLARIN (Ljubešić & Erjavec, 2016; Ljubešić et al., 2016). After automatic tagging, the subcorpora were manually checked and all tagging errors were corrected. Then, the additional phonological, morphological, and syntactic analyses of the Riddys corpus were performed separately for each subcorpus.

The grapheme analysis[3] included data on their overall amount and mean phonological length of words. All tokens of each subcorpus have been divided into syllables by means of publicly available syllabification algorithm, which is based on the maximum onset principle (Meštrović et al., 2015). Data on types of syllables that occur in a given subcorpus were obtained. In morphological analysis, tokens were systematised and processed according to types of speech, so data on lemmas and morphosyntactic tags are also available (Ljubešić et al., 2016).

Riddys corpus was also tagged syntactically, using the ReLDIanno tool, according to the formalism of the UD project (Agić and Ljubešić, 2015; Samardžić et al., 2017). From this marked corpus, data on the total number of sentences and fragments that are not proper sentences (e.g., enumerations) were extracted.

[3] Croatian language is a language with transparent orthography, which means that there is a high correspondence between phonemes and graphemes. In this paper we use the term grapheme for marking the smallest unit of written language, and phoneme if there is a reference to spoken language.

Conducted analyses

The multilevel analysis of the Riddys corpus described above enabled a detailed description of linguistic (primarily phonological, lexical and syntactic) characteristics of texts, i.e., the analysis of written texts to which lower-primary pupils are exposed.

Phonological characteristics of words in the texts were observed on the level of word length expressed in the number of graphemes.[4] Although there are still no fully reliable and systematic studies on the influence of a chosen unit of measurement on the analyses performed, nor on their interrelatedness, the choice of a unit may indirectly influence the author's conclusions and affect the distributional model of the observed text (see Grzybek, 2007 and Smith, 2012). This influence will vary in different orthographies, i.e., it will depend on the orthographic transparency and syllabic system of the observed language. Therefore, choosing an appropriate measure is extremely important when comparing texts written in different languages that may have different orthographies, and authors choose one or the other depending on the purpose of their analysis and the observed language. This study analyses texts written in a language with transparent orthography intended for children with different literacy levels, i.e., different levels of reading skills. Children attending the first grade, and to some extent those attending the second, are considered beginner readers. They usually decode a written text on the level of individual graphemes, while children in the third and fourth grade can grasp larger units (for example, syllables or words) during reading. According to the new curriculum for the Croatian language, learning uppercase and lowercase letters is expected at the end of the second grade of elementary school. In addition, the process of reading and writing automatisation – and entering the so-called orthographic phase, which refers to reading while relying on larger units (syllables and words; Firth, 1985) – starts later. Precisely because of the differences in the level of automatisation of decoding skill and the fact that 50 % of the materials analysed in this study are texts intended for children who decode at the grapheme level, it was decided to analyse word length in the number of graphemes. The distribution of all graphemes in written materials was analysed, as well.

The description of lexical characteristics of the Riddys corpus is based on the lemma/token ratio (LTR) and the distribution of particular types of speech, as well as the most frequent and least frequent words in each subcorpus. Tokens are all words that occur in written materials, while lemmas represent different words that occur in a text, whereby they are reduced to their basic form. LTR is one of the measures of lexical diversity (see Jelaska & Baričević, 2012; Kuvač & Palmović, 2007).

Syntactic characteristics of Riddys corpus were examined exclusively on the level of the mean length of a sentence, measured in the number of words. The mean length of an utterance (MLU) was first described as a measure of language development in

[4] This measure has been in use as a measure of text quality and individual writing style since the 19th century (Mendenhall, 1887). Apart from graphemes it can also be measured in the number of syllables (Elderton, 1949).

1973 (Brown, 1973) when it was observed as a ratio of morphemes per utterance. With time, it was recognised that the interpretation of this measure depends on the morphological system of a language, which resulted in discussions on its suitability in morphologically richer and more complex language systems such as Croatian (Kuvač & Palmović, 2007). Today, it is recommended that in such languages, a syntactic complexity measure should be observed using the mean length of utterances in words (see Kelić et al., 2012; Kuvač & Palmović, 2007).

All linguistic characteristics were observed with respect to educational level (four grades/subcorpora), and the data expressed in proportions at the level of descriptive statistics were later compared using a *t*-test for proportions. The first purpose of the comparisons corresponded to the first aim of the paper; to describe and compare the linguistic content of written materials to which students of different ages and educational levels are exposed. The second purpose of the analyses corresponded to the second aim of the paper, which was to discuss written materials in relation to the principles of *plain language*.

Results

Phonological analysis of the Riddys corpus (word length and grapheme distribution)

To examine the differences in the distribution of words of different lengths considering the educational level, we examined the length of all words in each subcorpus. The overall mean length of words in all four subcorpora is lower than 5 (first grade = 4.53; second grade = 4.49; third grade = 4.19, and fourth grade = 4.17). However, the distribution of words considering their length is quite diverse (Figure 1). In the Riddys corpus, most words have five to seven graphemes (around 35 %), followed by words with one to two and three to four graphemes (23 % to 26 %). The stability of these percentages can be explained by the fact that certain high-frequency words in Croatian that express relations and references mostly contain between two and four graphemes (for example, auxiliary verbs, pronouns, connectives, and prepositions), while some of the most frequent nouns in texts for lower grades usually have three to six or seven graphemes (for example, *dan [day], kraj [end], mama [mom], škola [school], čovjek [man], dijete [child]*, etc.; see 3.2. Lexical analysis of the Riddys Corpus). Words with eight to ten graphemes constitute 11.5 % to 13 % of all lexical units, and textbooks for the first and the second-grade pupils even contain words with up to 14 to 16 graphemes (Figure 1). Words of such length are present despite the well-known fact that in the earliest period of formal education, children have not yet fully automatised their reading skills and, therefore, have difficulty mastering such large units.

Pupils who attend the first and the second grade are only beginning to read, and they decipher on the grapheme level. That is why texts intended for this age should have significantly fewer long words than texts designed for older students. Therefore, further analysis aimed to determine whether there are significant differences in the

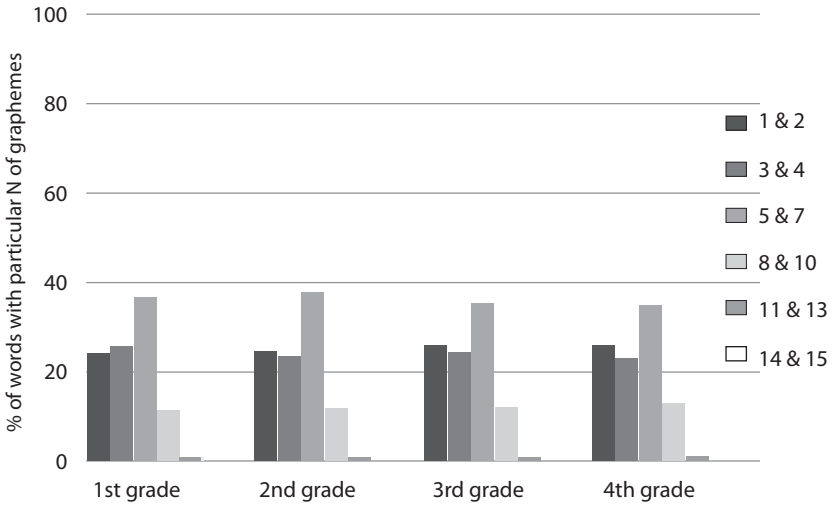

Figure 1. Percentage of words with different phonological length (measured in the number of graphemes) in written materials for the first four grades of elementary school.

distribution of words of different phonological lengths between the four analysed subcorpora. Due to the assumption that materials for the third and the fourth grade should have a larger number of long words, while those for the first and the second grade should have a greater number of shorter words, the differences in the distribution of words with a different number of graphemes will be presented.

A *t*-test showed that subcorpora of the first and the second grade do not differ significantly in the number of words with less than three and more than eight graphemes ($t<1.96$; $p>0.05$ for all comparisons), i.e., the written material in these textbooks is similar considering the amount of the shortest and the longest words. On the other hand, the first grade has a significantly greater proportion of words with three and four graphemes ($t=7.2125$; $p<0.00001$), while materials for the second grade have a higher number of words with five to seven graphemes ($t=-4.0427$; $p<0.00001$). Materials for the first and the third grade differ in the amount of longer words with eight to ten graphemes ($t=-3.7506$; $p=0.00018$), as well as words with 14 to 16 graphemes ($t=-2.3134$; $p=0.0208$), with third-grade textbooks consisting of a significantly higher number of longer words. On the other hand, the first grade has a larger amount of shorter and moderately long words (1 and 2 graphemes: $t=-8.5302$; $p<0.00001$; 3 and 4 graphemes: $t=6.4012$; $p<0.00001$; 5-7: $t=4.4786$; $p<0.00001$) compared to the third. There is no difference in the proportion of words containing 11 to 13 graphemes ($t<1.96$; $p>0.05$), i.e., words of that length are equally present in the textbooks for

these two grades. Comparisons of subcorpora of the first and the fourth grade showed significant differences in all comparisons; 1 and 2 graphemes (t=-10.3018; p<0.00001), 8–10 graphemes (*t*=-10.9623; *p*<0.0001), 11–13 graphemes (*t*=-5.8116; *p*<0.0001) and 14–16 graphemes (*t*=-5.101; *p*<0.0001) in favour of the fourth grade, and in some shorter and moderately long words with 3 and 4 (*t*=13.5566; *p*<0.00001) and 5–7 graphemes (*t*=7.0155; *p*<0.00001) in favour of the first.

The second and the third grade differ in the amount of shortest (1 and 2 graphemes: *t*=-7.8606; *p*<0.00001; 3 and 4 graphemes: *t*=-2.0717; *p*=0.0385) and longest words (11–13 graphemes: *t*=-14.269; *p*<0.0001; 14–16 graphemes: *t*=-2.5144; *p*=0.0121), both in favour of the third grade; while the second grade has a higher number of words with 5 to 7 graphemes (*t*=9.8578; *p*<0.00001). Materials for the second and the fourth grade differ significantly in all comparisons, whereby the second grade has a larger proportion of words with 3 and 4 and 5–7 graphemes (*t*=5.2474; *p*<0.00001 and *t*=12.8494; *p*<0.00001), while the fourth grade has a higher number of words with 1 and 2 graphemes (*t*=-9.7861; *p*<0.00001), 8–10 (*t*=-9.4512; *p*<0.0001), 11–13 (*t*=-5.5365; *p*<0.0001) and 14–16 graphemes (*t*=-5.6005; *p*<0.0001).

Finally, even the materials for the third and the fourth grade differ considering the distribution of the shortest (1 and 2 graphemes: *t*=-2.1283; *p*=0.033) and the longest words (8–10 graphemes: *t*=-10.2063; *p*<0.0001; 11–13 graphemes: *t*=13.1104; *p*<0.0001; 14–16 graphemes: *t*=-4.5879; *p*<0.0001) in favour of the higher grade. On the other hand, materials for the third grade have a significantly higher number of moderately long words (3 and 4 graphemes: *t*=9.6401; *p*<0,00001; 5–7 graphemes: *t*=3,3459; *p*<0.00001).

The conducted analyses show that written materials used in the first and the second grade consist of a higher number of words containing between three and seven graphemes, while materials used in the third and the fourth grade have a much higher distribution of longer words, as well as the shortest ones – those with 1 and 2 graphemes. The differences are the least observable between the first and the second grade, while they gradually become larger when entering the third grade. The most significant increase in word length occurs in the transition from third to fourth grade.

In order to examine the distribution of individual graphemes in each subcorpus, which is essential information for the order of grapheme learning in early reading and writing, the frequency of occurrence of all graphemes in written materials has been examined. In all grades, the most frequent graphemes are the vowels *a, i, e,* and *o,* followed by the consonants *n, r, s,* and *t* and the vowel *u* (with differences in the frequency of the graphemes *n* and *r*; *r* is more common than *n* in the second grade), while *dž, f* and *đ* are least frequent in all grades (Table 1).

Table 1
Grapheme distribution in the four subcorpora.

Grapheme	frequency (1st grade)	frequency (2nd grade)	frequency (3rd grade)	frequency (4th grade)	Total (1st-4th grade)
A	33100	40722	86352	115985	**276159**
I	27681	33974	69753	97636	**229044**
E	24364	30452	64803	86558	**206177**
O	23522	28470	60928	83517	**196437**
N	13339	16159	33123	46127	**108748**
R	12795	16456	32165	44475	**105891**
S	11692	15992	30614	42734	**101032**
U	12094	15391	30258	41812	**99555**
T	11605	15026	30011	42457	**99099**
J	11227	13663	28611	39840	**93341**
K	10451	12315	24442	34692	**81900**
M	10026	12170	24471	34983	**81650**
V	9193	11926	23446	32999	**77564**
D	8372	11271	23331	29847	**72821**
L	9145	10658	22427	29637	**71867**
P	8225	10124	18841	24994	**62184**
Z	4474	5908	11321	15382	**37085**
B	4457	5646	10432	14213	**34748**
G	3631	5107	10980	13431	**33149**
Š	4270	4613	9168	12136	**30187**
Č	3820	4294	8303	12161	**28578**
C	3333	4059	6912	9048	**23352**
Ć	2207	2735	6255	7971	**19168**
LJ	1911	2743	4915	7366	**16935**
Ž	1942	2364	3916	5988	**14210**
NJ	1570	2405	3927	5879	**13781**
H	1421	1896	3255	5114	**11686**
Đ	587	654	1236	1875	**4352**
F	445	429	494	612	**1980**
DŽ	131	113	80	127	**451**

Lexical analysis of the Riddys Corpus

As a lexical measure that provides insight into the lexical richness of the textbooks and other required reading materials for students, the aforementioned lemma/token ratio (LTR) has been calculated. When the ratio is higher (closer to value 1), lexical diversity is higher, i.e., there is a higher number of novel words and fewer words that repeat throughout the text. A lower ratio (closer to value 0) indicates lower lexical diversity or richness (see Kuvač & Palmović, 2007).

For the lexical analysis, we observed lemmas found at the ends of the continuum of frequency distribution – those that occur 12 or more times, which is a desirable amount in order to be learned in the written form, and those that occur only once. There were only 8.7 % to 12.6 % of lemmas with a satisfactory occurrence in all grades. On the other hand, in textbooks for the first and the second grade, more than 50 % of lemmas occurred only once, i.e., in one form and one context (51.2 % and 51.8 %), while in the third and the fourth grade, that percentage is 48.6 %, i.e., 47 % (Figure 2).

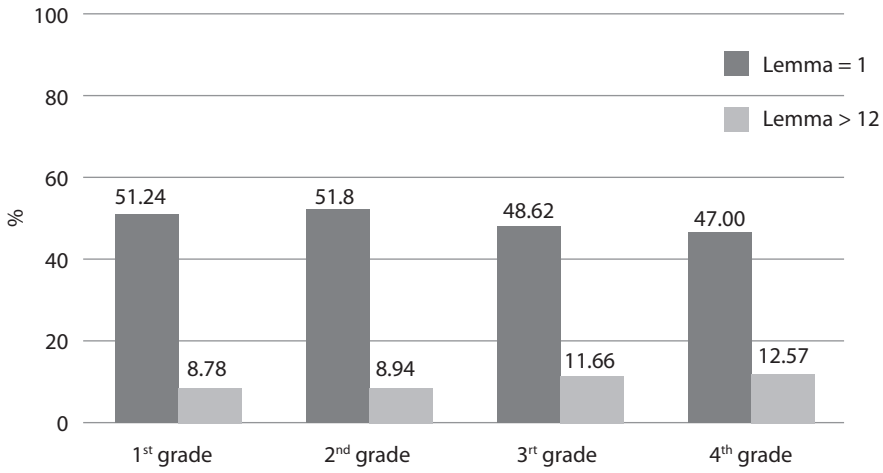

Figure 2. Percentage of lemmas in textbooks (1st–4th grade) that occur once and those that occur at least 12 times in different forms.

When written materials in all subcorpora are compared with respect to the number of lemmas that occur in the smallest amount (=1), only the materials in the first and the second grade do not differ from one another (t=-0.808; p=0.418). On the other hand, the materials in the first and the third grade (t=3.770; p=0.0002), the first and the fourth (t=6.308; p<0.00001), as well as the second and the third (t=4.904; p<0.00001), the second and the fourth (t=7.649; p<0.00001), and the third and the fourth grade (t=2.844; p=0.005) differ significantly considering the number of words that occur only once. This amount is always significantly higher in lower grades (the second in relation to the third and the fourth, and the third in relation to the fourth). Furthermore, the amount of lemmas with appropriate frequency (≥12) was compared. Once again, there

are no differences when comparing the subcorpora of the first and the second grade (t=-0.237; p=0.810), while the remaining differences are significant (the first and the third: t=-6.820; p<0.00001; the first and the fourth: t=-8.982; p<0.00001; the second and the third: t=-6.983; p<0.00001; the second and the fourth: t=-9.324; p<0.00001; the third and the fourth: t=-2.442; p=0.01468). In this case, the significance goes in favour of the higher grade, i.e., there is a significantly higher number of words with appropriate frequency in the third and the fourth grades in relation to the second, and in the fourth grade in relation to the third. This is also visible in percentages illustrated in Figure 2.

This unfavourable tendency is also confirmed in the lemma token ratio as a direct measure of lexical diversity of a text. Despite the observable increase in the number of tokens through grades, which is expected and logical, lemma token ratio is higher in the lower grades than in the third and the fourth (Table 2).

Table 2

The number of lemmas and tokens and lemma token ratio in written texts per grade.

	1st	2nd	3rd	4th
N lemma	8304	9949	14325	17520
N token	58375	71932	160039	212367
Lemma token ratio	0,14	0,14	0,09	0,08

For the purpose of further analyses, only lemmas with appropriate frequency have been singled out N(1st grade) = 729; N(2nd grade) = 889; N(3rd grade) = 1671; N(4th grade) = 2202), and for each of them a type of speech was determined, followed by their distribution in the Riddys corpus. As illustrated in Figure 3 and Table 3, in all grades, nouns (from 43 % to 46 %) and verbs (from 21 % to 25 %) are the highest in frequency, followed by adjectives (from 10 % to 14 %) and adverbs (around 6 %). Particles, interjections, and numbers have the lowest occurrence (0.3 to 2 %).

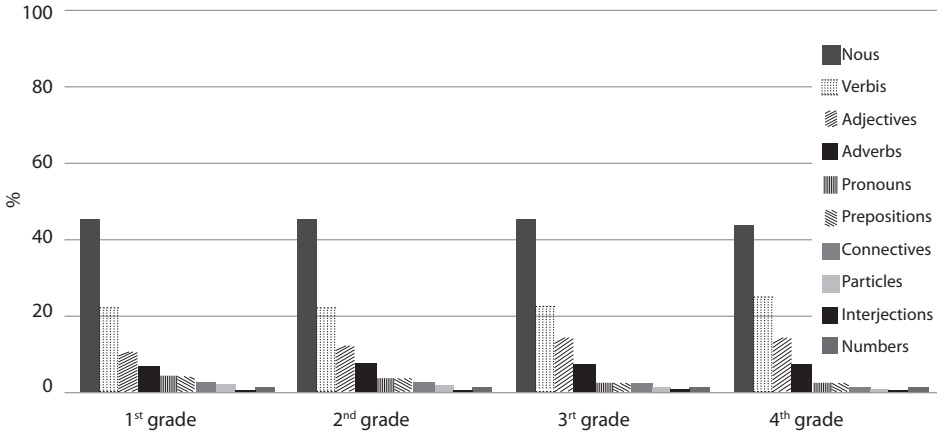

Figure 3. Distribution of all types of speech per grade.

The percentage of flective types of speech in all grades exceeds 83 %, while the percentage of inflective types of speech is 14 % in the first and the second grade, 12 % in the third and 10.5 % in the fourth grade (Table 3).

Table 3

Frequency and percentages of different types of speech in written materials in the four subcorpora (analysis included only lemmas with frequency ≥ 12).

CORPUS	1st grade: N (%)	2nd grade: N (%)	3rd grade: N (%)	4th grade: N (%)
N nouns	336 (46.09)	408 (45.89)	756 (45.24)	965 (43.82)
N verbs	163 (22.36)	190 (21.37)	374 (22.38)	551 (25.02)
N adjectives	77 (10.56)	106 (11.92)	234 (14)	319 (14.49)
N pronouns	31 (4.25)	32 (3.59)	40 (2.39)	41 (1.86)
N numbers	9 (1.23)	9 (1.01)	17 (1.02)	24 (1.09)
N FLECTIVE TYPES OF SPEECH (%)	**616 (84.49)**	**745 (83.80)**	**1421 (85.04)**	**1900 (86.29)**
N adverbs	48 (6.58)	62 (6.97)	114 (6.82)	148 (6.72)
N prepositions	28 (3.84)	32 (3.59)	42 (2.51)	36 (1.63)
N connectives	15 (2.06)	18 (2.02)	21 (1.26)	20 (0.91)
N particles	13 (1.78)	15 (1.68)	18 (1.08)	15 (0.68)
N interjections	2 (0.27)	2 (0.22)	6 (0.36)	11 (0.49)
N INFLECTIVE TYPES OF SPEECH (%)	**96 (14.54)**	**129 (14.51)**	**201 (12.03)**	**230 (10.45)**

In order to examine the significance of differences in the amount of flective and inflective types of speech in all written materials, again, a *t*-test for proportions was conducted. Comparisons did not reach significance, i.e., there are no differences in the amount of flective/inflective types of speech between the grades ($t<1.96$; $p>0.05$).

For every type of speech, the most frequent and the least frequent words were singled out. Out of the most prevalent types of speech in the Riddys corpus (nouns, verbs, adjectives and adverbs), the most frequent words that occur equally in written materials for all four grades are: škola*, dan, mama, riječ, dijete,* čovjek*, kuća, vrijeme, kraj [school, day, mother, word, child, man, house, time, end]* (nouns); *biti, htjeti, imati, moći, voljeti, ići, trebati, reći, znati, morati, pisati, doći [to be, want, have, can, love, go, need, say, know, must, write, come]* (verbs); *dobar, malen, nov, sav, velik, sam, star, crn, bijel, crven [good, small, new, all, big, alone, old, black, white, red]* (adjectives); *kako, kad, zašto, mnogo, gdje, tako, kada, jako, malo, sad [how, when, why, much/many, where, so, when, very, little, now]* (adverbs).

Syntactic analysis of the Riddys corpus

The length and complexity of a sentence are two characteristics of syntactic structures that can significantly influence their processing and comprehension. Even though a processing load of a sentence is reflected in its length and type, whereby a higher number of words most often corresponds to greater syntactic complexity (Kelić et al., 2012), simpler structures can also bear high processing costs (for example, a three-

part structure with an uncanonical word order; *object-predicate-subject*) (Babić, 1997). In this study, syntactic complexity is observed exclusively regarding length – namely, mean length of sentence in words (*MLSw*).

Mean length of sentence in the *School Corpus of Written Language* increases with the increase of educational level, i.e., in the first grade it equals to approximately five words, while in the fourth grade, it equals around seven (Figure 4). Pupils who attend the third grade of elementary school read the longest sentences (MLSw=7.71), which is when the highest leap in the increase of elements comprising one syntactic structure occurs (an increase in comparison to 5.69 in the second grade).

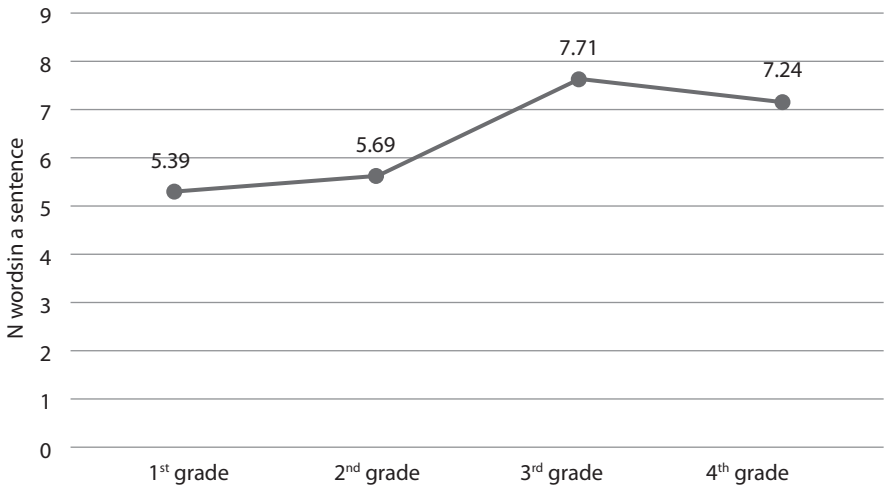

Figure 4. Overview of the mean length of sentence in the four subcorpora (1st-4th grade).

Discussion

The purpose of this paper was to determine whether the written materials on which the development of automatised reading and writing is based are, at present, more aligned with the principles of *plain language*, i.e., a language that is primarily based on cognitive and language abilities and needs of its users. The main assumptions were that the complexity of written materials will gradually increase regarding word length, lexical diversity and length of sentences.

Mean length of words is almost the same in all grades, but the distribution of words concerning their length is different. The analysis of word length measured in graphemes illustrates that within the Riddys corpus, most words consist of five to seven graphemes (around 30 %). They are followed by the shortest words with one and two and those with three and four graphemes (23 % to 26 %). In line with our expectations, the subcorpora of the third and the fourth grades consist of longer words in comparison to the subcorpora of the first two grades. Written materials for the first and the second grade, intended for the youngest pupils who are only starting the process of mastering

literacy, are the most similar. As pupils move to higher grades, school demands and the length of words in the texts they read also increase.

Nevertheless, the differences in length were not found in certain individual analyses; texts for the first and the third grade contain an equal number of words with 11, 12, and 13 graphemes. The most considerable discrepancies in word length emerge in the fourth grade of elementary school when materials become filled with significantly longer words compared to the materials intended for third-grade pupils. Furthermore, higher grades have higher proportions of the shortest words containing one and two graphemes. These are primarily function words whose quantity increases with the number of words in syntactic structures due to their role in establishing inter-sentential relations.

Grapheme distribution in the analysed materials indicates a very high consistency throughout the texts; a high frequency of vowels and the lowest frequency of graphemes that are generally less present in Croatian words (*dž, f, đ, h, nj*). The same distribution is visible in earlier analyses of phoneme distribution in spoken language conducted by Vuletić (1991), which was subsequently compared with grapheme distribution in written materials. Grapheme distribution substantially overlaps with the order of grapheme learning, which is one of the principles in designing primers (Bežen & Reberski, 2014; Lenček & Gligora, 2010). The study by Kuvač Kraljević, Lenček, and Matešić (2019), using factor analysis in naming uppercase letters in preschool children, identified two factors: the first encompassed naming of the so-called universal graphemes, which children learn first because they are more visually recognisable and relatively common in speech and writing (e.g., all vowels *a, e, i, o, u* and some consonants *b, k, l, m, n, r, s, t*), and the second defined naming seven graphemes characteristic for the Croatian Latin script, i.e., graphemes with diacritical marks (***č, ć, đ, ž***) and digraphs (*dž, lj, nj*). These two factors correspond to the first and the last graphemes considering their distribution in the written materials analysed here.

In computational and corpus linguistics, numerous linguistic measures were developed as objective indicators of the complexity of spoken or written content (for example, the measure of lexical diversity and syntactic complexity; see Crystal, 1979; MacWhinney, 2000). These measures are used to describe the degree of language acquisition in different populations, such as bilingual speakers or speakers with language disorders (e.g., Hržica et al., 2020; Kelić et al., 2012; Lu, 2011), but they can provide the data on the suitability of linguistic content for the population for whom it is intended. The latter approach was applied for the linguistic analyses in this paper, using one of the lexical measures as an indicator of lexical diversity (lemma token ratio) and one syntactic measure (mean length of sentence) as an indicator of syntactic complexity of written materials intended for young students in the lower grades of elementary school.

The analysis on the lexical level was conducted by extracting the lemmas that occur twelve or more times and those that occur only once. The main reason for this is the fact that children learn words and completely grasp their meaning only when they are

exposed to them in different written contexts and forms at least twelve times (Beck et al., 2002). The analysis showed that the proportion of lemmas occurring only once is the highest in the first and second grades. This means that children who only begin to read are frequently required to decode words phonologically (i.e., to decipher one grapheme at a time), which is a more time-consuming and cognitively demanding way of reading. Contrary to this, a more frequent occurrence of the same lemma speeds up the decoding process on a whole-word level because frequent exposure to the same lemma enables such a form to be decoded orthographically. The proportion of lemmas which occur 12 or more times is the highest in the third and the fourth grade. The obtained data are not in line with the principles of *plain language*. *Plain language* should be based on the words that are frequently used in the spoken language, that are semantically unambiguous, and whose meaning is additionally reinforced by their use in different contexts.

Furthermore, the lemma token ratio as a measure for lexical diversity does not go in favour of the first and second-grade students, where this ratio seems to be higher. This points to the greater occurrence of novel words, i.e., illustrates that there are fewer repetitions and a lesser amount of different forms of the same lemma (Kuvač & Palmović, 2007). Even though different textbooks were analysed for the purpose of this study, the fact is that first-grade pupils encounter a significant number of written texts for the first time, from which they must acquire new knowledge. This process would be easier and faster if they were exposed to the same word multiple times and in different contexts. It seems that texts in primers, besides stimulating reading, always aim at lexical enrichment. Undeniably, these two dimensions – reading and lexicon – are closely related and interdependent. However, it is unlikely that during the period in which a child should master this rather demanding phonological-orthographic skill, strong lexical development will also occur. Such demands are especially overwhelming when a child comes into contact with a particular phonological structure only once.

The analysis of the distribution of different types of speech shows the expected dominance of content/lexical words in relation to function words. The order of the obtained occurrences is in line with the data on the ratio of content and function words from Cvikić (2002), although the percentages in the two papers slightly differ. This can result from the differences in the content between the analysed textbooks. Cvikić (2002) based her analysis on two primers, while this study includes written materials from several school subjects intended for pupils in all four lower-primary years.

The analysis of the mean length of sentence in words (MLSw) showed that the third-grade children read the longest sentences. This is when the highest increase of elements forming a syntactic structure occurs (7.71 compared to 5.69 in the second grade). The data on sentence length obtained in this study reflect, to some extent, the data on language production of children entering elementary school and children attending the lower elementary school grades. Namely, Kelić et al. (2012) showed that the mean length of sentence in narrative samples of six-year-olds on average amounts

to 5.76, and that of nine-year-olds around 7.29. According to the principles of *plain language*, syntactic structures are too demanding to process if a reader or a listener cannot retain them in short-term memory due to their unfavourable length. Based on the conducted analyses, it can be concluded that the analysed structures in textbooks gradually increase in length as their users progress in education.

Although this is one of the most comprehensive studies of written materials in textbooks for lower grades conducted to date, there are certain limitations, so the conclusions should be taken with a grain of salt. First, the corpus consists of relatively few textbooks, which limits the possibility of generalisations. Moreover, the study is limited in the scope of linguistic analyses and the use of measures, especially on the syntactic level. The complexity of syntactic structures, apart from their length, refers to their type. Therefore, in order to draw more concrete conclusions, the types of syntactic structures that occur in the textbook materials for lower grades should also be analysed.

Conclusion

Plain language is based on the principle of simplicity, directness and concreteness. In order to apply it in written materials intended for children in elementary school, it is essential to choose appropriate words, i.e., those that occur in ' 'children's spoken language, those that have the appropriate phonological length, words commonly present in written materials, and words embedded in syntactic structures of the appropriate length. It is assumed that the main reason for designing overly demanding texts lies precisely in the neglect of the language abilities with which a beginner reader enters the decoding process to finally be able to understand a written text.

The analysis of randomly chosen written materials to which pupils are exposed in the first four grades of elementary school implies that the principles of *plain language* are fulfilled on the phonological and syntactic levels. However, the latter was measured only through the length of syntactic structures. Word length gradually increases with the chronological and educational age of a student, i.e., with the transition to a higher grade. Moreover, there is an overlap in the distribution of graphemes in ' 'children's textbooks, and the order of their mastering in the preschool period. This means that the graphemes to which children are exposed the most in written materials are the graphemes they learn first. The progression in the increase of length of syntactic structures indicates a ' 'pupil's gradual exposure to structures with a higher number of words.

It seems that out of three observed language components – phonology, lexicon, and syntax – only lexicon does not follow the principles of *plain language*. Overrepresentation of words that occur only once in the written materials for the first two grades of elementary school does not contribute to faster and more efficient development of reading skills. Constant exposure to novel and unfamiliar words with a new phonological plan poses high demands on the ' 'child's phonological processing. Exposing children to

words that have been stored in their mental lexicon based on spoken communication, but more than 12 times in different contexts, improves not only their phonological decoding skills but also facilitates their successful shift to orthographic decoding, i.e., the process of automatisation of reading. This ensures a faster mapping of a word with its meaning.

The authors of written materials for first-grade pupils, most likely encouraged by a theoretical determinant on the link between lexicon and reading, seem to focus more on unfamiliar words for which an automatic retrieval of meaning and their storage in the child's mental lexicon is implied. Unfortunately, these processes do not necessarily occur in parallel, i.e., simultaneously. Reading is a highly complex phonological skill that requires prior linguistic knowledge and a synchronised activity of several cognitive mechanisms, sometimes resulting in a slower enrichment of language with new elements.

When clear learning objectives are set that match the child's language abilities, it will be easier to achieve full compliance of the written materials for pupils in grades one through four with the principles of *plain language*. *Plain language* will ensure that children who have just begun to read can easily find what they need, understand what they find, and apply the newly acquired knowledge successfully.

Acknowledgment

This paper is a part of a project *Development of an innovative diagnostic instrument for early recognition of children with dyslexia (RiDDys) (KK.01.2.1.02.0167),* within the call Increasing the development of new products and services arising from R&D activities – Phase II (IRI 2).

References

Agić, Ž., Ljubešić, N. (2015). Universal Dependencies for Croatian (that work for Serbian, too). In: *The 5th Workshop on Balto-Slavic Natural Language Processing*, 1-8.

Arias-Badia, B., Matamala, A. (2020). Audio description meets Easy-to-Read and Plain Language: results from a questionnaire and a focus group in Catalonia. *Zeitschrift* für *Katalanistik, 33,* 251-270.

Babić, Z. (1997). Utjecaj različitih rečeničnih struktura na jezičnu obradu. In: Ljubešić, M. (Ed.), *Jezične teškoće školske djece: oblici, uzroci, posljedice, otklanjanje* (pp. 153-176). Zagreb: Školska knjiga.

Beck, I. L., McKeown, M. G., Kucan, L. (2002). *Bringing words to life: Robust Vocabulary Instruction.* New York, London: The Guilford Press.

Bežen, A., Reberski, S. (2014). *Početno pisanje na hrvatskome jeziku: priručnik uz "Hrvatski pravopis".* Zagreb: Institut za hrvatski jezik i jezikoslovlje.

Brown, R. (1973). *A first language: The early stages.* Cambridge, MA: Harvard University Press. https://doi.org/10.4159/harvard.9780674732469

Celik, B. (2020). A Study on the Factors Affecting Reading and Reading Habits of Preschool Children. *International Journal of English Linguistics, 10*(1), 101-114. https://doi.org/10.5539/ijel.v10n1p101

Crystal, D. (1979). *Working with LARSP.* London: Edward Arnold.

Cvikić, L. (2002). Pretpostavljeno i očekivano znanje prvašića. In: Vodopija, I. (Ed.), *Zbornik radova s međunarodnog stručnog i znanstvenog skupa u europskoj godini jezika - Dijete i jezik danas* (pp. 55-72). Osijek: Sveučilište J. J. Strossmayera.

Elderton, W. P. (1949). A few statistics on the length of English words. *Journal of the Royal Statistical Society. Series A (General), 112*(4), 436-445. https://doi.org/10.2307/2980766

Frith, U. (1985). Beneath the surface of developmental dyslexia. In: Patterson, K. E., Marshall, J. C., Coltheart, M. (Ed.), *Surface Dyslexia: Neuropsychological and Cognitive Studies of Phonological Reading (1st ed.)* (pp. 301-330). London: Routledge. https://doi.org/10.4324/9781315108346-18

Grzybek, P. (2007). History and methodology of word length studies. In: Grzybek, P. (Ed.), *Contributions to the Science of Text and Language* (pp. 15-90). Dordrecht: Springer. https://doi.org/10.1007/1-4020-4068-7_2

Hržica, G., Košutar, S., Kramarić, M. (2019). Rječnička raznolikost pisanih tekstova osoba s razvojnim jezičnim poremećajem (Lexical Diversity in Written Texts of Persons with Developmental Language Disorder). *Hrvatska revija za rehabilitacijska istraživanja [Croatian Review for Rehabilitation Research], 55*(2), 14-30. https://doi.org/10.31299/hrri.55.2.2

International Plain Language Federation (IPLF). *Plain Language Definitions*. Retrieved from https://www.iplfederation.org/plain-language/

Jelaska, Z., Baričević, V. (2012). Leksička jednostavnost i značenjska složenost Ivanova evanđelja (Simplicity and complexity of John's Gospel vocabulary). *LAHOR - Članci i rasprave, 13*, 102-137.

Kelić, M., Hržica, G., Kuvač Kraljević, J. (2012). Mjere jezičnog razvoja kao pokazatelji posebnih jezičnih teškoća (Measurements of Language Development as Markers of Specific Language Impairment (SLI)). *Hrvatska revija za rehabilitacijska istraživanja [Croatian Review for Rehabilitation Research], 48*(2), 23-40.

Kovačević, M., Kuvač, J. (2004). Jezik udžbenika i jezik djeteta: razumiju li se oni? In: Bacalja, R. (Ed.), *Zbornik radova sa znanstveno-stručnog skupa s međunarodnom suradnjom Dijete, odgojitelj i učitelj Stručni odjel za izobrazbu učitelja i odgojitelja predškolske djece*. (pp. 93-100). Sveučilište u Zadru: Zadar.

Kuvač, J., Palmović, M. (2007). *Metodologija istraživanja dječjega jezika.* Jastrebarsko: Naklada Slap.

Kuvač Kraljević, J., Lenček, M. (2020). Školski *korpus pisanog jezika (School Corpus of Written Language) (Riddys).* Zagreb: Edukacijsko-rehabilitacijski fakultet.[5]

Kuvač Kraljević, J., Lenček, M., Matešić, K. (2019). Phonological awareness and letter knowledge: indicators of early literacy in Croatian. *Hrvatski časopis za odgoj i obrazovanje [Croatian Journal of Education],* 21(4), 1263-1293.

[5] Korpus je razvijen za potrebe projekta i nije javno dostupan.

Lenček, M., Gligora, J. (2010). U početku bijaše riječ: o početnicama i čitanju (In the Beginning Was The Word: On Primers and Reading). *Logopedija, 2*(1), 36-44.

Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL 'writers' language development. *TESOL QUARTERLY, 45*(1), 36-62. https://doi.org/10.5054/tq.2011.240859

Ljubešić, N., Erjavec, T. (2016). Corpus vs. lexicon supervision in morphosyntactic tagging: the case of Slovene. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 1527-1531.

Ljubešić, N., Klubička, F., Agić, Ž., Jazbec, I. P. (2016). New inflectional lexicons and training corpora for improved morphosyntactic annotation of Croatian and Serbian. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 4264-4270.

Maaß, C. (2020). *Easy Language – Plain Language – Easy Language Plus: Balancing Comprehensibility and Acceptability*. Berlin: Frank & Timme GmbH. https://doi.org/10.26530/20.500.12657/42089

MacWhinney, B. (2000). *The CHILDES Project: Tools for analysing talk. 3rd Edition*. Mahwah, NJ: Lawrence Erlbaum Associates.

McGeown, S. P., Osborne, C., Warhurst, A., Norgate, R., Duncan, L. G. (2016). Understanding ' 'children's reading activities: Reading motivation, skill and child characteristics as predictors. *Journal of research in reading*, *39*, 109-125. https://doi.org/10.1111/1467-9817.12060

Mendenhall, T. C. (1887). The characteristic curves of composition. *Science, 9*(214), 237-249. https://doi.org/10.1126/science.ns-9.214S.237

Meštrović, A., Martinčić-Ipšić, S., Matešić, M. (2015). Postupak automatskoga slogovanja temeljem načela najvećega pristupa i statistika slogova za hrvatski jezik (Syllabification based on maximal onset principle for Croatian). *Govor [Speech]*, *32*(1), 3-34.

Miličić, S. (2016). *Leksička i sintaktička analiza početnica*. Diplomski rad, Sveučilište u Zagrebu, Edukacijsko-rehabilitacijski fakultet.

Miličić, S., Kuvač Kraljević, J., Lenček, M. (2017). Leksička analiza početnica (Lexical Analysis of Primers). *Logopedija, 7*(1), 30-37. https://doi.org/10.31299/log.7.1.5

Montolío, E. (2019). Lingüística, comunicación y transferencia de conocimiento a la sociedad: Un reto para el siglo XXI, *Palimpsesto 15*, 54-67.

Morgan, P. L., Fuchs, D. (2007). Is there a bidirectional relationship between children's reading skills and reading motivation?. *Exceptional children, 73*(2), 165-183. https://doi.org/10.1177/001440290707300203

Nemeth-Jajić, J. (2007). *Udžbenici hrvatskoga jezika u razrednoj nastavi (Textbooks of Croatian Language in the Elementary School).* Redak: Split.

Official Gazette no. 27/10. and 55/11. Ordinance on the Textbook Standard and Members of Expert Committees for the Evaluation of Textbooks and other Educational Materials (2013). http://public.mzos.hr/Default.aspx?art=12470

Official Gazette no. 9/2019. Ordinance on the Textbook Standard and Members of Expert Committees for the Evaluation of Textbooks and other Educational Materials (2019). https://narodne-novine.nn.hr/clanci/sluzbeni/2019_01_9_196.html

Radić, Ž., Kuvač Kraljević, J., Kovačević, M. (2010). Udžbenik kao poticaj ili prepreka leksičkom razvoju (A textbook as facilitating or hindering factor of lexical development). *Lahor:* časopis *za hrvatski kao materinski, drugi i strani jezik, 9*, 43-59.

Radić Tatar, I. (2013). Ovladanost vrstama rečenica na kraju predškolske dobi (Acquisition of Sentence Types at the End of Preschool Age). *Lahor:* časopis *za hrvatski kao materinski, drugi i strani jezik, 2*(16), 165-188.

Samardžić, T., Starović, M., Agić, Ž., Ljubešić, N. (2017). Universal dependencies for Serbian in comparison with Croatian and Other Slavic Languages. In: *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing (BSNLP 2017),* Valencia, Spain. https://doi.org/10.18653/v1/W17-1407

Scott, J. A. (2005). Creating Opportunities to Acquire New Word Meanings From Text. In: Hiebert, E. E., Kamil, M. L. (Ed.), T*eaching and Learning Vocabulary: Bringing Research to Practice* (pp. 69-94). New Jersey: Lawrence Erlbaum Associate.

Smith, R. D. (2012). Distinct word length frequencies: distributions and symbol entropies. *Glottometrics 23*, 7-22.

Vuletić, D. (1991). *Istraživanje govora.* Zagreb: Fakultet za defektologiju Sveučilišta u Zagrebu.

Ana Matić Škorić
University of Zagreb, Faculty of Education and Rehabilitation Sciences, Department of Speech and Language Pathology
University Campus Borongaj, Borongajska cesta 83f,
10000 Zagreb, Croatia
ana.matic@erf.unizg.hr

Jelena Kuvač Kraljević
University of Zagreb, Faculty of Education and Rehabilitation Sciences, Department of Speech and Language Pathology
University Campus Borongaj, Borongajska cesta 83f,
10000 Zagreb, Croatia
jelena.kuvac@erf.unizg.hr

Mirjana Lenček
University of Zagreb, Faculty of Education and Rehabilitation Sciences, Department of Speech and Language Pathology
University Campus Borongaj, Borongajska cesta 83f,
10000 Zagreb, Croatia
mirjana.lencek@erf.unizg.hr

Appendix

Textbooks and required readings included in the Riddys corpus

1) Croatian Language:

Pčelica 1; part 1 and 2
Pčelica 1; workbook
Authors: Veronek Germadnik, Velić, Križman Roškar
Publisher: Profil

Čitam i pišem 2
Authors: Pavličević, Velički, Domišljanović
Publisher: Alfa

Hrvatski jezik 3 + workbook
Authors: Pavličević, Domišljanović
Publisher: Alfa

Čarolija riječi
Authors: Težak, Polak, Cindrić
Publisher: Alfa

Integrated textbook:
Zlatna vrata: integrated textbook for Croatian Language and Literacy in the 2nd grade
Authors: Krmpotić, Ivić
Publisher: Školska knjiga

2) Science:

Pogled u svijet 1: Tragom prirode i društva
Authors: Škreblin, Svoboda Arnautov, Basta
Publisher: Profil

Eureka! 3
Authors: Bakarić Palička, Ćorić
Publisher: Školska knjiga

Naš svijet 4
Authors: Kisovar Ivanda, Letina, Nejašmić, De Zan, Vranješ Šoljan
Publisher: Školska knjiga

3) Religious Education:

Rastimo u zahvalnosti
Authors: Jakšić, Mićanović
Publisher: Glas Koncila

4) Mathematics:

Matematičkim stazama 4
Authors: Paić, Manzoni, Kosak, Marjanović
Publisher: Školska knjiga

Required readings:

1st grade:
Plesna haljina žutog maslačka, Sunčana Škrinjarić

2nd grade:
Grga čvarak, Ratko Zvrko
Pismo iz Zelengrada, Nevenka Videk

3rd and 4th grade:
Mate Lovrak: Vlak u snijegu
Ivana Brlić Mažuranić: Šegrt Hlapić
Sanja Pilić: Nemam vremena
Anto Gardaš: Duh u močvari
Sanja Polak: Dnevnik Pauline P.
Zvonimir Balog: Ja magarac
Nikola Pulić: Ključić oko vrata

+ several volumes (issues) of the periodicals: Radost, Smib and Modra lasta

Jasan jezik i pisana udžbenička građa nižih razreda: jesu li usklađeni?

Sažetak

Jasan jezik (engl. plain language) koncept je koji se prije pola stoljeća proširio Europom, a odnosi se na komunikaciju temeljenu na iskazima i strukturi iz koje slušatelji i čitatelji mogu lako pronaći i razumjeti potrebnu informaciju. U ovom istraživanju koncept jasnoga jezika predstavljen je u kontekstu ranoga obrazovanja, analizom pisane građe kojoj su izloženi učenici nižih razreda osnovne škole. Cilj je ovoga rada analizirati fonološka, leksička i sintaktička obilježja udžbenika u prvom, drugom, trećem i četvrtom razredu te dobivene podatke povezati s općim načelima jasnoga jezika. Za potrebe rada razvijena je baza tekstova – Školski korpus pisanoga jezika (Riddys), s ukupno 502 713 pojavnica. Analizirana je zastupljenost grafema, vrsta riječi, kao i duljina riječi i rečenica. Rezultati pokazuju kako od tri promatrane jezične sastavnice samo rječnik ne slijedi načela jasnoga jezika. Prevelika zastupljenost riječi koje se pojavljuju tek jednom u cijeloj pisanoj udžbeničkoj građi u prvim dvama razredima osnovne škole nikako ne pridonosi bržem i učinkovitijem ovladavanju vještinom čitanja.

Ključne riječi: *jasan jezik; jezična analiza udžbenika; početno čitanje i pisanje; razredna nastava; Školski korpus pisanoga jezika (Riddys).*

Uvod

Jasan jezik

Prije 60 godina Europom se proširila krilatica *Jezik za sve* kojom se željelo upozoriti na teškoće na koje prosječni govornik nekog jezika svakodnevno nailazi u razumijevanju različitih vrsta tekstova i poruka (Maaß, 2020).

Kao odgovor na prepoznate teškoće tijekom 70-ih godina prošloga stoljeća uveden je *jasan jezik* (engl. *plain language*) kao koncept i instrument kojim se željelo ne samo riješiti navedene teškoće, nego i osvijestiti važnost promišljanja o odgovarajućem načinu primjene jezika (Arias-Badia i Matamala, 2020). Kao pandan pojmu *jasan jezik* navodi se i pojam *jasna komunikacija* jer se, prema nekim autorima poput Montolío (2019), njime još šire zahvaća opseg ljudske interakcije.

Prema International Plain Language Federation (IPLF) *komunikacija jasnim jezikom je ona u kojoj su iskazi, struktura i oblik toliko jasni da čitatelji kojima je namijenjena*

mogu lako pronaći ono što im treba, razumjeti ono što pronađu i upotrijebiti te informacije (IPLF, 2021). Glavni je cilj *jasnoga jezika* učiniti sadržaj razumljivim onima kojima je on i namijenjen pa stoga krilatica *manje je više* najbolje opisuje njegovu bit. *Jasan jezik* ne koristi se samo za pisani jezik, tj. tekst, već i za govoreni jezik promoviran u javnome, političkome, pravnome, obrazovnome i medijskome prostoru (Arias-Badia i Matamala, 2020).

Najčešći uzrok nedovoljnoj razumljivosti teksta ili poruke jest to što jezik teksta nije prilagođen jezičnomu znanju čitatelja i/ili slušatelja. Da bi tekst bio razumljiv, potrebno je integrirati nekoliko načela koja će to osigurati, primjerice analitičnost, izravnost i konkretnost u pisanju, točnost i sustavnost u iznošenju sadržaja te naglašavanje bitnoga (Maaß, 2020). To podrazumijeva odabir učestalih riječi nedvosmislenoga značenja. Odabir sintaktičkih struktura mora biti takav da one svojom duljinom mogu biti zadržane u kratkoročnome pamćenju slušatelja, a njihova obrada koncizna i brza. Oblikovanje teksta pak mora odgovarati spoznajnoj razini slušatelja, a tema bi trebala biti takva da pobuđuje daljnji interes i osigurava dublje i kritičnije promišljanje o njemu.

Smatra se da su najčešći razlozi neprimjerene uporabe jezika upravo autorovo nerazumijevanje spoznajnih, a posebice jezičnih mogućnosti i potreba njegovih čitatelja i korisnika. Stoga, ako se želi oblikovati poruka koja će uspješno doći do korisnika i pritom biti jednako razumljiva svima, nužno je kontinuirano promišljati o tome što čitatelj može i zna. U ovom istraživanju koncept *jasnoga jezika* bit će predstavljen u kontekstu ranoga obrazovanja, a problematizirat će se kroz analizu pisane udžbeničke građe kojoj su izloženi učenici u nižim razredima osnovne škole.[1]

Jezik udžbenika

Djeca koja tek ulaze u školski sustav značajno se razlikuju s obzirom na količinu vremena tijekom kojega su bila izložena pisanome jeziku i s obzirom na način na koji ih se njemu izlagalo. Razlog tomu leži u nizu unutarnjih i vanjskih čimbenika kao što su obitelj i obiteljska pismenost, uključenost djeteta u vrtić ili druge formalizirane predškolske programe te opće spoznajne i psihološke osobine djeteta (Celik, 2020). Polazak u prvi razred svoj djeci osigurava kontinuiranu i svakodnevnu izloženost pisanomu materijalu i stabilno okruženje za daljnji jezični razvoj. Međutim, brzina i uspješnost kojom će učenici razumijevati prve pisane riječi u školskome sustavu ponovno su određene njihovim spoznajnim i psihološkim osobinama, predvještinama čitanja i pisanja, metodama poučavanja, ali i obilježjima pisanih materijala na kojima se ta poduka temelji.

Iako je u Udžbeničkom standardu (2013; 2019) jasno navedeno da su tekstovi koje Ministarstvo znanosti i obrazovanja RH (MZO) potvrđuje i odobrava za uporabu primjereni djeci s obzirom na njihov spoznajni i jezični razvoj te da na taj način

[1] Koncept jasnoga jezika, osim jezične strukture pisane i govorene komunikacije, definira i načela grafičkoga prikaza i vizualnih aspekata dizajna dokumenata (vidi više u Adrias-Badia i Matamala, 2020), no te komponente pisane građe nisu u interesu ovoga rada, zbog čega neće biti analizirane.

potiču kreativnost i razvoj kritičkoga razmišljanja i metakognicije, niz istraživanja usmjerenih na izučavanje sadržaja i jezika početnica i udžbenika pokazuje suprotno. Na nezadovoljavajuću usklađenost prvotnih udžbenika s djetetovim jezičnim sposobnostima i vještinama pismenosti upozorava se od početka 2000. godina, kada se pojavljuju prve analize početnica i kada započinje proučavanje obilježja jezika kojemu su djeca izložena (primjerice, Cvikić, 2002; Kovačević i Kuvač, 2004; Lenček i Gligora, 2010; Radić, Kuvač Kraljević i Kovačević, 2010), ali i kada se pojavljuju vrlo opsežne analize pojedinih udžbenika u razrednoj nastavi (primjerice, udžbenika Hrvatskoga jezika; v. Nemeth-Jajić, 2007). Pritom se posebno ističe neprikladnost u odabiru riječi i sintaktičkih struktura koje nadilaze djetetovo semantičko znanje i mogućnost jezične obrade.

Tri su preduvjeta nužna za poticanje rječničkoga razvoja pomoću čitanja: razvijena vještina dekodiranja grafema (veze grafem-fonem); mogućnost prepoznavanja nepoznatih riječi; mogućnost zaključivanja iz konteksta (Beck, McKeown i Kucan, 2002). Nedovoljna uspješnost u bilo kojem od ovih preduvjeta negativno će utjecati na daljnji rječnički, odnosno ukupni jezični razvoj, a posljedično i smanjiti motivaciju za daljnje izlaganje pisanomu tekstu (McGeown i sur., 2016). Naime, između čitanja, odnosno ovladanosti ovom vještinom i motivacijom za čitanje postoji dvosmjerna veza, što se u procesu obrazovanja ne smije zanemariti (Morgan i Fuchs, 2007). Kako bi se motivacija očuvala, a vještine prikladno razvile, sadržaji kojima su djeca izložena trebali bi odražavati njihov jezični razvoj. Prema Scott (2005), rječnički se razvoj može potaknuti ako su početni čitatelji učestalo izloženi poznatim riječima, podjednako u ponavljajućim i u novim kontekstima. Beck i suradnici (2002) navode kako čitatelj nepoznatoj riječi treba biti izložen i do dvanaest puta kako bi u potpunosti ovladao njezinim značenjem. Prevelik omjer nepoznatih riječi u odnosu na poznate nepovoljno će utjecati na bogaćenje rječnika, a što je još problematičnije, i na samo ovladavanje čitanjem. Pisani jezik djeci treba biti sintaktički jasan, a riječi prikladne njihovoj dobi i primjenjive u različitim kontekstima. Istraživanja koja su dosad provedena u hrvatskome jeziku pokazala su da tomu često nije tako. Kovačević i Kuvač (2004) ustanovile su da se značajan dio riječi u početnicama – čak više od 50 % – pojavljuje samo jednom. Nadalje, podatci o tome koliko su čitanke ujednačene, odnosno koliko jezičnoga sadržaja dijele, upućuju na samo 30 % preklapanja u rječničkome sadržaju, pri čemu su podudaranja veća u broju nepunoznačnih nego u broju punoznačnih riječi (primjerice, imenica i glagola) (Cvikić, 2002). Istraživanja koja su uslijedila (Lenček i Gligora, 2010; Radić i sur., 2010) iznjedrila su podatke gotovo identične onima s početka 2000. Jedno od recentnijih istraživanja rječničkoga sadržaja početnica (Miličić, Kuvač Kraljević i Lenček, 2017) uključuje tri nasumično odabrane početnice.[2]

[2] U svakoj školskoj godini broj odobrenih početnica je različit, u nekim školskim godinama broj je veći od sedam.

To je istraživanje ponovno potvrdilo koliko su početnice zahtjevne i neprimjerene, kako s obzirom na pojavnost riječi (novih i nepoznatih) tako i s obzirom na udio niskočestotnih riječi. Građa kojoj su djeca izložena stoga niti je u skladu s obrazovnim standardima, niti s nalazima psiholingvističkih istraživanja o obilježjima jezičnoga razvoja i tijeka opismenjavanja.

U školskoj se dobi pod utjecajem jezičnoga razvoja i izlaganja pisanomu tekstu, osim rječničkih znanja proširuju i sintaktički oblici (duljinom i vrstom). I dok se u hrvatskoj literaturi pronalaze podatci o zastupljenosti vrsta rečenica u dječjoj proizvodnji (Radić Tatar, 2013) ili o njihovoj zastupljenosti u početnicama hrvatskoga jezika (Miličić, 2016), nažalost nedostaju detaljni podatci o duljini tih struktura. Kovačević i Kuvač (2004) upozoravaju na prisutnost izrazito dugih rečenica (s više od 40 riječi) u prvim dječjim udžbenicima koje, uz zasićenje riječima čije je značenje učenicima nepoznato, pridonose nerazumijevanju teksta. Vuletić (1991) je analizom govorenoga korpusa pokazala da se izgovorena rečenica sastoji u prosjeku od 6,45 riječi pri čemu su najučestalije rečenice s četiri ili pet riječi.

Prethodno navedeni podatci o leksičkoj i sintaktičkoj strukturi pisane građe proizlaze isključivo iz analiza građe korištene u prvome razredu osnovne škole koji su bili materijalom obrazovne pouke tijekom proteklih 20 godina. Svrha je ovoga rada istražiti i usporediti pisanu udžbeničku građu svih nižih razreda – od prvoga do četvrtoga – na kojoj se temelji razvoj početnoga i automatiziranoga čitanja i pisanj, te utvrditi je li ona danas usklađenija s načelima *jasnoga jezika* čije planiranje polazi od razumijevanja spoznajnih i jezičnih mogućnosti i potreba čitatelja.

Cilj rada

Cilj je istraživanja analizirati fonološka, leksička i sintaktička obilježja udžbenika nižih razreda osnovne škole te usporediti zastupljenost pojedinih elemenata s navedenih triju jezičnih sastavnica u prvome, drugome, trećemu i četvrtome razredu. Nadalje, cilj je rada dobivene podatke o obilježjima pisane udžbeničke građe povezati s općim načelima *jasnoga jezika* te raspraviti koliko analizirani udžbenici zrcale ta načela, poput zadovoljavajuće učestalosti riječi i duljine sintaktičkih struktura. Pretpostavlja se da će se pisana građa kojoj su izloženi učenici u nižim razredima osnovne škole usložnjavati s porastom obrazovnoga stupnja na odabranim mjerama: duljina riječi mjerena brojem grafema, rječnička raznolikost i duljina rečenica.

Metode rada

Analizirana građa: Školski korpus pisanoga jezika (Riddys)

Za potrebe ovoga rada razvijena je baza udžbeničkih tekstova namijenjenih polaznicima nižih razreda osnovne škole iz četiriju nastavnih predmeta: Hrvatski jezik, Matematika, Priroda i društvo te Vjeronauk. Osim pisane građe iz navedenih predmeta, uključeno je i približno 20 % lektirne građe onih naslova koji se preklapaju u različitim popisima te nekoliko brojeva dječjih obrazovnih časopisa koje je odobrilo

MZO i koji se nabavljaju preko škole. Sva je građa prepisana i objedinjena u Školski korpus pisanog*a jezika* (*Riddys;* Kuvač Kraljević i Lenček, 2020; nastao za potrebe ciljeva projekta Riddys). Popis udžbenika, autora i izdavača naveden je u Dodatku.

Tekstovi koji su zajedno imali 502 713 pojavnica (i oko 45 000 rečenica) podijeljeni su u četiri podkorpusa prema razredima. Sva su četiri podkorpusa lematizirana te morfosintaktički i sintaktički obilježena alatom ReLDIanno koji je javno dostupan u sklopu slovenskoga čvora projekta CLARIN (Ljubešić i Erjavec, 2016; Ljubešić i sur., 2016). Nakon automatskoga označavanja, podkorpusi su ručno pregledani i ispravljene su pogreške u označavanju. Potom se pristupilo dodatnoj fonološkoj, morfološkoj i sintaktičkoj analizi korpusa *Riddys*, zasebno za svaki podkorpus.

Analiza grafema[3] uključivala je podatke o njihovom ukupnome broju i prosječnoj fonološkoj duljini riječi. Određena je i fonemska struktura svakoga podkorpusa, odnosno raspodjela riječi prema broju grafema u riječima. Sve su pojavnice svakoga podkorpusa slogovane pomoću javno dostupnoga algoritma za slogovanje koji se temelji na načelu najvećega pristupa (Meštrović, Martinčić-Ipšić i Matešić, 2015) i dobiveni su podatci o pojavnosti slogova u zadanim podkorpusima. U morfološkoj su analizi pojavnice sistematizirane i frekvencijski obrađene prema vrstama riječi, tako da su dostupni podatci i po lemama i po morfosintaktičkim oznakama (Ljubešić i sur., 2016).

Korpus *Riddys* je uporabom alata ReLDIanno označen i sintaktički, i to prema formalizmu zadanom UD projektom (Agić i Ljubešić, 2015; Samardžić i sur., 2017). Iz tako označenoga korpusa izdvojeni su podatci o ukupnome broju rečenica te odsječaka koji nisu potpuni sintaktički iskazi (primjerice, nabrajanja).

Provedene analize

Opisana višerazinska obrada korpusa *Riddys* omogućila je detaljan opis lingvističkih (prvenstveno fonoloških, leksičkih i sintaktičkih) obilježja tekstova, odnosno analizu pisanoga jezika kojemu su izloženi polaznici nižih razreda osnovne škole.

Fonološka obilježja riječi u tekstovima promatrana su na razini duljine riječi izražene brojem grafema.[4] Iako još uvijek ne postoje sasvim pouzdana i sustavna istraživanja o utjecaju odabrane mjerne jedinice na provedene analize, kao ni o njihovim međusobnim odnosima, odabir jedinice neizravno može utjecati na autorove zaključke te na prikaz distribucijskoga modela promatranoga teksta (v. Grzybek, 2007 i Smith, 2012). Taj će utjecaj biti drugačiji u različitim pismima, odnosno ovisit će o transparentnosti ortografije i slogovnomu sustavu promatranoga jezika. Odabir prikladne mjere stoga je posebno važan pri uspoređivanju tekstova pisanih na različitim jezicima koji mogu imati različite ortografske sustave, a autori će se prikloniti jednoj ili drugoj mjeri što

[3] Hrvatski je jezik transparentne ortografije, što znači da postoji visoka podudarnost između fonema i grafema. U ovom će se radu ponajviše rabiti pojam grafem koji označava najmanju jedinicu pisanoga jezika. Pojem fonem rabit će se samo u kontekstu govorenoga jezika.

[4] Ova se mjera kao mjera kvalitete teksta i individualnoga stila pisanja upotrebljava još od 19. stoljeća (Mendenhall, 1887), a osim u grafemima može se mjeriti i u slogovima (Elderton, 1949).

prvenstveno ovisno o svrsi analiza i jeziku koji promatraju. U ovome se radu analiziraju tekstovi pisani jezikom transparentne ortografije kojima su izloženi učenici različitoga stupnja pismenosti, tj. ovladanosti čitalačkom vještinom. Učenici prvoga i do neke mjere učenici drugoga razreda smatraju se početnim čitateljima koji tekst najčešće i većinom dekodiraju na razini pojedinačnih grafema, dok polaznici trećih i četvrtih razreda prilikom čitanja zahvaćaju veće cjeline (primjerice, slogove i/ili riječi). Prema novome kurikulu za Hrvatski jezik, učenje velikih i malih formalnih (tiskanih) i pisanih grafema očekuje se tek na kraju drugoga razreda osnovne škole. Osim toga, proces automatizacije čitanja i pisanja i ulaženje u tzv. ortografsku fazu koja podrazumijeva čitanje s oslanjanjem na veće jedinice (slogove i riječi; Frith, 1985) započinje kasnije. Upravo zbog razlika u stupnju automatizacije vještine dekodiranja i činjenice da 50 % građe koja se u ovome radu promatra čine tekstovi kojima su izloženi učenici koji dekodiraju na razini pojedinačnih grafema, duljina riječi promatra se u broju grafema. Analizirana je i zastupljenost svih grafema u pisanoj građi.

Opis leksičkih obilježja *Riddys* korpusa temelji se na omjeru natuknica i pojavnica (ONP) te prikazu distribucije pojedinih vrsta riječi, kao i najučestalijih i najmanje učestalih riječi u svakome od podkorpusa. Pojavnice su sve riječi koje se pojavljuju u nekoj pisanoj građi, dok natuknice predstavljaju različite riječi koje se pojavljuju u tekstu, pri čemu su svedene na svoj osnovni oblik. ONP jedna je od mjera rječničke raznolikosti teksta (Jelaska i Baričević, 2012; Kuvač i Palmović, 2007).

Sintaktička se obilježja *Riddys* korpusa promatraju isključivo na razini prosječne duljine rečenice (PDR) mjerene u broju riječi. Mjera prosječne duljine iskaza (PDI) prvi je put kao mjera jezičnoga razvoja opisana 1973. godine (Brown, 1973), kada se promatrala kao omjer morfema po iskazu. S vremenom se ustanovilo da interpretacija te mjere ovisi o morfološkome sustavu jezika, zbog čega su uslijedile rasprave o njezinoj prikladnosti u morfološki bogatijim i složenijim jezičnim sustavima kao što je hrvatski (Kuvač i Palmović, 2007). Danas se u takvim jezicima mjeru složenosti na razini sintakse preporučuje promatrati mjerom prosječne duljine iskaza u riječima (v. Kelić, Hržica i Kuvač Kraljević, 2012; Kuvač i Palmović, 2007).

Sva su lingvistička obilježja promatrana s obzirom na obrazovni stupanj (četiri razreda/podkorpusa), a podatci koji su izraženi u proporcijama na razini deskriptivne statistike naposljetku su uspoređeni uporabom *t* testa za proporcije. Prva je svrha usporedbi usklađena s prvim ciljem rada; opisati i usporediti jezične sadržaje kojima su izloženi učenici različite dobi i obrazovnoga stupnja. Druga je svrha provedenih analiza usklađena s drugim ciljem rada koji je usmjeren na raspravu o usklađenosti pisane udžbeničke građe s načelima *jasnoga jezika*.

Rezultati

Fonološka analiza korpusa Riddys (duljina riječi i zastupljenost grafema)

Kako bi se ispitale razlike u odnosima zastupljenosti riječi različite duljine s obzirom na obrazovni stupanj, promatrala se duljina svih pojavnica u pojedinim podkorpusima.

Njihova ukupna prosječna duljina u podkorpusima svih razreda manja je od 5 (prosječna duljina riječi u podkorpusu prvoga razreda jest 4,53, drugoga 4,49, trećega 4,19, a četvrtoga 4,17). Ipak, distribucija riječi s obzirom na duljinu vrlo je raznolika (Slika 1). Ukupno je u korpusu *Riddys* najviše riječi s pet do sedam grafema (oko 35 %), nakon čega slijede riječi s jednim i dva, odnosno tri i četiri grafema (23 % do 26 %). Stabilnost ovih udjela može se pripisati činjenici da neke visoko učestale riječi u hrvatskome jeziku kojima se izražavaju odnosi i referencija sadrže između dva i četiri grafema (primjerice, pomoćni glagoli, zamjenice, veznici i prijedlozi), a neke od najzastupljenijih imenica u tekstovima nižih razreda nerijetko imaju tri do šest ili sedam grafema (primjerice *dan, kraj, mama, škola, čovjek, dijete* i sl.; v. *Leksička analiza korpusa Riddys*). Riječi s osam do 10 grafema imaju pojavnost između 11,5 % i 13 %, a čak i udžbenici namijenjeni učenicima prvoga i drugoga razreda sadrže riječi s 14 do 16 grafema (Slika 1), unatoč spoznajama o nedovoljno automatiziranoj vještini čitanja u najranijemu razdoblju formalnoga obrazovanja i težini zahvaćanja ovakvih cjelina.

Slika 1

Polaznici prvoga i drugoga razreda još uvijek su početni čitatelji koji dekodiraju na razini grafema, stoga bi tekstovi namijenjeni ovoj dobi trebali imati značajno manji udio duljih riječi u odnosu na tekstove namijenjene starijim učenicima. U daljnjoj se analizi stoga nastojalo utvrditi postoje li značajne razlike u udjelu riječi različite fonološke duljine između četiri analizirana podkorpusa. Zbog pretpostavke da će udio duljih riječi biti veći u trećemu i četvrtome razredu, a udio kraćih u prvome i drugome razredu, prikazat će se razlike u udjelima riječi s različitim brojem grafema.

Provedbom *t* testa utvrđeno je da se podkorpusi prvoga i drugoga razreda ne razlikuju značajno u udjelima riječi s manje od tri i više od osam grafema ($t < 1,96$; $p > 0,05$ za sve usporedbe), odnosno udžbenička građa ovih dvaju razreda slična je s obzirom na udio najkraćih i najduljih riječi. Nasuprot tomu, prvi razred ima značajno veći udio riječi s tri i četiri grafema ($t = 7,2125$; $p < 0,00001$), dok je građa drugoga razreda bogatija riječima s pet do sedam grafema ($t = -4,0427$; $p < 0,00001$). Građa prvoga i trećega razreda razlikuje se u udjelu duljih riječi s osam do 10 grafema ($t = -3,7506$; $p = 0,00018$), kao i riječi s 14 do 16 grafema ($t = -2,3134$; $p = 0,0208$), pri čemu udžbenici trećega razreda imaju značajno više duljih riječi. S druge strane, prvi razred prednjači u udjelu kraćih i srednje dugih riječi (1 i 2 grafema: $t = -8,5302$; $p < 0,00001$; 3 i 4 grafema: $t = 6,4012$; $p < 0,00001$; 5-7 grafema: $t = 4,4786$; $p < 0,00001$) u odnosu na treći razred. Nema razlike u udjelu riječi s 11 do 13 grafema ($t < 1,96$; $p > 0,05$), odnosno riječi te duljine u jednakoj su mjeri zastupljene u udžbenicima ovih dvaju razreda. Usporedbe podkorpusa prvoga i četvrtoga razreda iznjedrile su značajne razlike u svim usporedbama; 1 i 2 grafema ($t = -10,3018$; $p < 0,00001$), 8 - 10 grafema ($t = -10,9623$; $p < 0,0001$), 11 - 13 grafema ($t = -5,8116$; $p < 0,0001$) te 14 - 16 grafema ($t = -5,101$; $p < 0,0001$) u korist četvrtoga razreda, a nešto kraćih i umjereno

dugih riječi s 3 i 4 (t = 13,5566; p < 0,00001) i 5 - 7 grafema (t = 7,0155; p < 0,00001) u korist prvoga razreda.

Drugi i treći razred razlikuju se s obzirom na udio najkraćih (1 i 2 grafema: t = -7.8606; p < 0,00001; 3 i 4 grafema: t= -2,0717; p = 0,0385) i najduljih riječi (11 - 13 grafema: t = -14,269; p < 0,0001; 14 - 16 grafema: t = -2,5144; p = 0,0121), oba u korist trećega razreda; dok drugi razred ima veći udio riječi s 5 do 7 grafema (t = 9,8578; p < 0,00001). Građa drugoga i četvrtoga razreda značajno se razlikuje u svim usporedbama, pri čemu drugi razred prednjači u udjelu riječi s 3 i 4 te 5 - 7 grafema (t = 5,2474; p < 0,00001 i t = 12,8494; p < 0,00001), dok četvrti razred ima veći udio riječi s 1 i 2 grafema (t = -9,7861; p < 0,00001), 8 - 10 (t = -9,4512; p < 0,0001), 11 - 13 (t= -5,5365; p < 0,0001) te 14 - 16 grafema (t = -5,6005; p < 0,0001).

Naposljetku, čak se i građa trećega i četvrtoga razreda razlikuje s obzirom na distribuciju najkraćih (1 i 2 grafema: t = -2,1283; p = 0,033) i najduljih riječi (8 - 10 grafema: t = -10,2063; p < 0,0001; 11 - 13 grafema: t = 13,1104; p < 0,0001; 14 - 16 grafema: t = -4,5879; p < 0,0001) u korist višega razreda. S druge strane, građa trećega razreda ima značajno veći udio riječi srednje duljine (3 i 4 grafema: t = 9,6401; p < 0,00001; 5 - 7 grafema: t = 3,3459; p < 0,00001).

Iz provedenih je analiza razvidno da pisana udžbenička građa prvoga i drugoga razreda ima veći udio riječi fonološke duljine između tri i sedam grafema, dok građa trećega i četvrtoga razreda obiluje duljim riječima, ali i najkraćim riječima s jednim i dva grafema. Najmanje su razlike između prvoga i drugoga razreda, razlike s prelaskom u treći razred postupno postaju veće, dok najveći porast duljine riječi nastupa na prijelazu iz trećega u četvrti razred.

Kako bi se utvrdila zastupljenost pojedinačnih grafema u svakom od podkorpusa, a što predstavlja važan podatak za poredak učenja grafema u početnome čitanju i pisanju, promatrana je učestalost pojavljivanja svih grafema u pisanoj građi. U svim su razredima najučestaliji grafemi vokali *a, i, e, o,* nakon čega slijede konsonanti *n, r, s, t* i vokal *u* (s razlikama u frekvenciji grafema *n* i *r*; *r* je učestaliji od *n* isključivo u drugom razredu), dok su *dž, f* i đ najmanje učestali u svim razredima (Tablica 1).

Tablica 1

Leksička analiza korpusa Riddys

Kao mjera koja pruža uvid u rječničko bogatstvo udžbenika i lektirnih sadržaja kojima su izloženi polaznici nižih razreda osnovne škole uzet je ranije opisan omjer natuknica i pojavnica (ONP). Što je taj omjer veći (bliži vrijednosti 1) to je veća rječnička raznolikost, odnosno više je novih riječi i manje je riječi koje se ponavljaju. S druge strane, što je taj omjer manji (bliži vrijednosti 0) to je rječnička raznolikost manja (v. Kuvač i Palmović, 2007).

U leksičkoj analizi izdvojene su natuknice koje se nalaze na krajnostima kontinuuma učestalosti – one koje se u različitim oblicima pojavljuju dvanaest i više puta, a što je poželjno kako bi se naučile u pisanom obliku te one koje se pojavljuju samo jednom.

Natuknica sa zadovoljavajućom pojavnošću u svim je razredima bilo samo 8,7 % do 12,6 %. S druge strane, u udžbenicima za prvi i drugi razred više od 50 % natuknica pojavilo se samo jednom, tj. u jednome obliku i kontekstu (51,2 % i 51,8 %), dok za treći i četvrti razred taj postotak iznosi 48,6 %, odnosno 47 % (Slika 2).

Slika 2

Kada se pisana građa svih podkorpusa međusobno usporedi s obzirom na udio natuknica s najmanjom pojavnošću (= 1), građa prvoga i drugoga razreda jedine se međusobno ne razlikuju (t = -0,808; p = 0,418). S druge strane, građa prvoga i trećega (t = 3,770; p = 0,0002), prvoga i četvrtoga (t = 6,308; p < 0,00001), kao i građa drugoga i trećega (t = 4,904; p < 0,00001), drugoga i četvrtoga (t = 7,649; p < 0,00001) te trećega i četvrtoga razreda (t = 2,844; p = 0,005) značajno se razlikuju s obzirom na udio natuknica koje se pojavljuju samo jednom. Taj je udio uvijek značajno veći u nižem razredu (drugom u odnosu na treći i četvrti te trećem u odnosu na četvrti). Jednako tako, uspoređen je i udio natuknica s prikladnom pojavnošću (≥ 12). Ponovno nema razlika kad se uspoređuju podkorpusi prvoga i drugoga razreda (t = -0,237; p = 0,810), dok su preostale razlike značajne (prvi i treći: t = -6,820; p < 0,00001; prvi i četvrti: t = -8,982; p < 0,00001; drugi i treći: t = -6,983; p < 0,00001; drugi i četvrti: t = -9,324; p < 0,00001; treći i četvrti: t = -2,442; p = 0,01468). Smjer značajnosti u ovom je slučaju u korist višega razreda, odnosno značajno je više riječi s prikladnom pojavnošću u trećemu i četvrtome razredu u odnosu na drugi, odnosno u četvrtome razredu u odnosu na treći. To je vidljivo i u postotnim udjelima prikazanima grafički na Slici 2.

Ova se nepovoljna tendencija potvrđuje i u omjeru natuknica i pojavnica kao izravnoj mjeri rječničke raznolikosti teksta. Unatoč uočenom porastu broja pojavnica kroz razrede, što je očekivano i logično, omjer natuknica i pojavnica veći je u nižim razredima negoli u trećemu i četvrtome razredu (Tablica 2).

Tablica 2

Za potrebe daljnjih analiza izdvojene su natuknice sa zadovoljavajućom pojavnošću: N(1. razred) = 729; N(2. razred) = 889; N(3. razred) = 1671; N(4. razred) = 2202) te je za svaku određeno kojoj vrsti riječi pripada, a potom i njihova distribucija u korpusu *Riddys*. Kako je prikazano na Slici 3 i u Tablici 3, u svim je razredima najveći udio imenica (od 43 % do 46 %) i glagola (od 21 % do 25 %), nakon čega slijede pridjevi (od 10 % do 14 %) i prilozi (oko 6 %). Najmanji je udio čestica, uzvika i brojeva (0,3 do 2 %).

Slika 3

Postotni udio promjenjivih vrsta riječi u svim razredima premašuje 83 %, dok udio nepromjenjivih vrsta riječi iznosi 14 % u prvome i drugome razredu, 12 % u trećemu te 10,5 % u četvrtome razredu (Tablica 3).

Tablica 3

Kako bi se ispitala značajnost razlika u udjelima promjenjivih i nepromjenjivih vrsta riječi u pisanoj građi četiriju razreda, ponovno je proveden t test za proporcije. Nijedna

usporedba nije pokazala značajne razlike, odnosno među razredima ne postoji značajna razlika u udjelu promjenjivih, odnosno nepromjenjivih vrsta riječi ($t < 1,96$; $p > 0,05$).

Za svaku su vrstu riječi izdvojene najučestalije i najmanje učestale riječi. Od vrsta riječi koje se u najvećem obimu pojavljuju u korpusu Riddys (imenice, glagoli, pridjevi i prilozi) najučestalije riječi, a ujedno i one koje se podjednako pojavljuju u pisanoj građi svih četiriju razreda, jesu: škola, dan, mama, riječ, dijete, čovjek, kuća, vrijeme, kraj (imenice); *biti, htjeti, imati, moći, voljeti, ići, trebati, reći, znati, morati, pisati, doći* (glagoli); *dobar, malen, nov, sav, velik, sam, star, crn, bijel, crven* (pridjevi); *kako, kad, zašto, mnogo, gdje, tako, kada, jako, malo, sad* (prilozi).

Sintaktička analiza korpusa Riddys

Duljina i složenost dva su obilježja sintaktičkih struktura koja mogu značajno utjecati na njihovu obradu i razumijevanje. Iako se obradbena složenost rečenice ogleda u njezinoj duljini i vrsti, pri čemu veći broj riječi najčešće zahtijeva i veću sintaktičku složenost (Kelić i sur., 2012), ipak i neke jednostavne strukture mogu imati visoke obradbene zahtjeve (primjerice, tročlana struktura *objekt-predikat-subjekt*) (Babić, 1997). Sintaktička je složenost u ovome radu izražena isključivo putem mjere duljine, i to prosječne duljine rečenice u riječima (PDR).

Prosječna duljina rečenica (PDR) u Školskom korpusu pisanog*a jezika* raste s porastom obrazovne dobi, odnosno u prvome razredu iznosi u prosjeku pet riječi, a u četvrtome oko sedam (Slika 4). Najdulje rečenice čitaju polaznici trećega razreda osnovne škole (PDR = 7,71), a tada se događa i najveći skok u porastu elemenata koji čine jednu sintaktičku strukturu (porast u odnosu na 5,69 u drugome razredu).

Slika 4

Diskusija

Svrha je ovoga rada bila istražiti je li pisana građa na kojoj se temelji razvoj početnoga i automatiziranoga čitanja i pisanja danas više usklađena s načelima *jasnoga jezika*, odnosno onoga jezika koji prvenstveno polazi od spoznajnih i jezičnih mogućnosti i potreba njegovih korisnika. Glavne su pretpostavke bile da će složenost pisane građe kojoj su izloženi učenici nižih razreda osnovne škole postupno rasti s obzirom na duljinu riječi, rječničku raznolikost i duljinu rečenica.

Prosječna duljina riječi u svim je razredima približno jednaka, ali je distribucija riječi s obzirom na njihovu duljinu drugačija. Analiza duljine riječi mjerene u grafemima pokazuje kako je u cijelom korpusu *Riddys* najviše riječi s pet do sedam grafema (oko 35 %). Slijede najkraće riječi s jednim do dva, odnosno one s tri do četiri grafema (23 % do 26 %). Sukladno očekivanjima, u podkorpusima trećega i četvrtoga razreda veća je zastupljenost duljih riječi u odnosu na podkorpuse prvih dvaju razreda. Pisana je građa prvoga i drugoga razreda, namijenjena najmlađim učenicima koji tek započinju proces opismenjavanja, u najvećoj mjeri ujednačena. Kako učenici prelaze u više razrede, tako rastu i školski zahtjevi, a sukladno tomu i duljina riječi

u tekstovima koje čitaju. Ipak, u nekim pojedinačnim analizama nisu pronađene razlike s obzirom na duljinu, pa se tako u tekstovima za prvi i treći razred u jednakoj mjeri čitaju riječi koje imaju čak i 11, 12 i 13 grafema. Najveće razlike u duljini riječi nastupaju s prelaskom u četvrti razred osnovne škole, kada udžbenička građa obiluje riječima koje su značajno dulje u odnosu na riječi koje su zastupljene u tekstovima namijenjenima polaznicima trećega razreda. Nadalje, u višim je razredima i veći udio najkraćih riječi s jednim i dvama grafemima. To su upravo funkcionalne riječi čija količina raste s porastom broja riječi u sintaktičkim strukturama zbog njihove uloge u uspostavljanju međurečeničnih odnosa.

Zastupljenost grafema u analiziranoj pisanoj građi pokazuje iznimno visoku ujednačenost kroz sve tekstove; visoku učestalost vokala te najmanju učestalost grafema koji su općenito rijetko zastupljeni u riječima hrvatskoga jezika (*dž, f, đ, h, nj*). Ista se distribucija pojavljuje i u ranijim analizama zastupljenosti fonema u govorenome jeziku koje je provela Vuletić (1991) i koju je naknadno usporedila s raspodjelom grafema u pisanoj građi. Zastupljenost grafema u pisanoj građi dobrim se dijelom preklapa i s redoslijedom učenja grafema, a što jest jedno od načela pri oblikovanju početnica (Bežen i Reberski, 2014; Lenček i Gligora, 2010). U istraživanju Kuvač Kraljević, Lenček i Matešić (2019) faktorskom su analizom na zadatku imenovanja velikih formalnih grafema predškolske djece izdvojena dva faktora: prvi faktor obuhvatio je imenovanje takozvanih univerzalnih grafema kojima djeca ovladavaju najranije jer se lako vizualno prepoznaju i učestali su u govoru i pismu (primjerice, svi vokali *a, e, i, o, u* te neki konsonanti, *b, k, l, m, n, r, s, t*) te drugi faktor koji je definiran imenovanjem sedam grafema osobitih za hrvatsku latinicu, odnosno grafema s dijakritičkim znakovima (č, ć, đ, ž) te digrafa (*dž, lj, nj*). Navedeni faktori odgovaraju prvim i posljednjim grafemima s obzirom na zastupljenost u ovdje analiziranoj pisanoj građi.

U području računalne i korpusne lingvistike razvijene su brojne jezične mjere kao objektivni pokazatelji složenosti govorenoga ili pisanoga sadržaja (primjerice, mjere rječničke raznolikosti i sintaktičke složenosti; v. Crystal, 1979; MacWhinney, 2000). Te se mjere rabe za opisivanje stupnja usvojenosti jezika kod različitih populacija kao što su dvojezični govornici ili govornici s jezičnim poremećajima (npr. Hržica, Košutar i Kramarić, 2020; Kelić i sur., 2012; Lu, 2011), ali mogu dati podatke o prikladnosti kakvoga jezičnoga sadržaja za populaciju kojoj je on namijenjen. U ovome se radu jezičnoj analizi pristupilo na potonji način, uporabom jedne od leksičkih mjera kao pokazatelja rječničke raznolikosti (omjer natuknica i pojavnica) te jedne sintaktičke mjere (prosječna duljina rečenice) kao pokazatelja sintaktičke složenosti pisane udžbeničke građe kojoj su izloženi polaznici nižih razreda osnovne škole.

Analiza na leksičkoj razini provedena je tako da su prvo izdvojene one natuknice koje se pojavljuju dvanaest i više puta te one koje se pojavljuju samo jednom. Temeljni je razlog tomu činjenica da djeca ovladavaju riječima i u potpunosti nauče njihovo značenje tek nakon što su joj u različitim pisanim kontekstima i oblicima izložena barem dvanaest puta (Beck i sur., 2002). Provedena analiza pokazala je da je broj natuknica

koje se pojavljuju samo jednom u cijeloj pisanoj građi najveći u prvome i drugome razredu. To znači da su početni čitatelji učestalo primorani fonološki dekodirati riječi (tj. čitati ih grafem po grafem) što je vremenski i obradbeno zahtjevniji način čitanja. Suprotno tomu, učestalije pojavljivanje istih natuknica (u različitim oblicima) u pisanoj građi ubrzava proces dekodiranja na razini cijele riječi jer se čestom izloženošću istoj natuknici takav oblik počinje dekodirati ortografskim putem. Zastupljenost natuknica čija je pojavnost veća od 12 najveća je u trećemu i četvrtome razredu. Dobiveni rezultati nisu u skladu s načelima *jasnoga jezika.* Naime, *jasan jezik* trebao bi se temeljiti na riječima koje su učestale u govorenoj uporabi i značenjski jednoznačne, a značenje se dodatno učvršćuje uporabom riječi u različitim kontekstima.

Nadalje, ni omjer natuknica i pojavnica kao mjera rječničke raznolikosti ne ide u prilog učenicima u prvome i drugome razredu u kojima je vrijednost toga omjera veća. To upućuje na veći broj novih riječi, odnosno pokazuje da je u pisanoj građi manje ponavljanja natuknica u različitim oblicima (Kuvač i Palmović, 2007). Iako je za potrebe ovoga rada analizirano nekoliko različitih udžbenika, činjenica je da se učenici u prvome razredu prvi put susreću s velikom količinom tekstova iz kojih trebaju steći nova znanja, a taj bi proces bio kudikamo lakši i brži da su više puta i u različitim kontekstima izložena istim riječima. Čini se kako se tekstovima u početnicama, uz učenje čitanja, stalno pokušava usporedno raditi i na bogaćenju rječnika. Iako je neupitno da su te dvije mjere - čitanje i rječnik - visoko povezane i međusobno ovisne, nije izgledno da u razdoblju u kojem dijete treba ovladati izuzetno zahtjevnom fonološko-ortografskom vještinom, ona bude jednako popraćena i snažnom razvojem rječnika. Takvi su zahtjevi posebice opterećujući kada se dijete određenoj fonološkoj strukturi izloži tek jednom.

Analiza distribucije s obzirom na vrstu riječi upućuje na očekivanu dominantnost punoznačnih/leksičkih riječi u odnosu na funkcionalne. Redoslijed ovih pojavnosti u skladu je s podatcima o omjeru punoznačnih i nepunoznačnih riječi autorice Cvikić (2002), iako se postotni udjeli u dvama radovima donekle razlikuju. To može biti posljedica sadržajnih razlika u vrstama analiziranih udžbenika. Naime, Cvikić (2002) je svoju analizu temeljila na dvjema čitankama za prvi razred, a ovo istraživanje uključuje pisanu građu drugih školskih predmeta namijenjenu učenicima svih nižih razreda.

Analiza prosječne duljine rečenica po broju riječi (PDR) pokazala je kako učenici trećega razreda čitaju najdulje rečenice. Tada se događa i najveći skok u porastu elemenata koji čine jednu sintaktičku strukturu (7,71 u odnosu na 5,69 u drugome razredu). Podatci o duljini rečenice dobiveni u ovom istraživanju donekle odražavaju podatke o jezičnoj proizvodnji djece školskih obveznika i polaznika nižih razreda osnovne škole. Naime, Kelić i suradnice (2012) pokazale su da prilikom pripovijedanja PDR šestogodišnjaka iznosi u prosjeku 5,76, a devetogodišnjaka 7,29. Prema načelima *jasnoga jezika* sintaktičke strukture obradbeno su prezahtjevne ako ih zbog neprikladne duljine čitatelj ili slušatelj ne može zadržati u kratkoročnom pamćenju. Na temelju provedenih analiza može se zaključiti da strukture zastupljene u analiziranoj udžbeničkoj građi postupno rastu duljinom kako njihovi korisnici napreduju u obrazovanju.

Iako je ovo jedno od najobuhvatnijih istraživanja pisane građe udžbenika nižih razreda, zbog nekih ograničenja njegove spoznaje ipak treba uzeti s oprezom. Prije svega, građu korpusa čini relativno oskudan broj udžbenika, zbog čega je i generalizacija otežana. Nadalje, istraživanje je ograničeno i opsegom jezičnih analiza i uporabom jezičnih mjera, posebice na sintaktičkoj razini. Složenost struktura, osim duljine, čini i vrsta rečenica, stoga bi za konkretnije zaključke trebalo analizirati i vrste sintaktičkih struktura koje se pojavljuju u pisanoj udžbeničkoj građi nižih razreda.

Zaključak

Jasan jezik temelji se na načelu jednostavnosti, izravnosti i konkretnosti. Da bi se on ostvario u pisanoj građi kojoj su učenici izloženi polaskom u školu, u prvome je redu nužno odabrati prikladne riječi, tj. one koje su već ovjerene u dječjemu govorenom jeziku, koje su odgovarajuće fonološke duljine, učestalo prisutne u pisanoj građi i ugrađene u sintaktičke strukture primjerene duljine. Smatra se da su najčešći razlozi oblikovanja prezahtjevnih tekstova upravo zanemarivanje jezičnih znanja s kojima čitatelj ulazi u proces dekodiranja s konačnim ciljem razumijevanja pisanoga teksta.

Provedena analiza nasumično odabrane pisane građe kojoj su učenici izloženi u prvim četirima razredima osnovne škole upućuje na to da su načela *jasnoga jezika* zadovoljena na fonološkoj i sintaktičkoj razini koja je u ovome radu mjerna samo putem duljine sintaktičkih struktura. Duljina riječi u pisanoj građi postupno raste s porastom kronološke i obrazovne dobi učenika, tj. prelaskom u viši razred. Štoviše, postoji i podudarnost u zastupljenosti grafema u udžbenicima i redoslijedu njihova poznavanja u predškolskom razdoblju, što znači da su djeca u pisanoj građi najčešće izložena onim grafemima koje prve i uče. Postupan porast duljine sintaktičkih struktura upućuje i na postupno izlaganje učenika strukturama s većim brojem riječi.

Od tri promatrane jezične sastavnice – fonologije, rječnika i sintakse – samo rječnik ne slijedi načela *jasnoga jezika*. Prevelika zastupljenost riječi čija je pojavnost u cijeloj pisanoj građi tek jednom u prvim dvama razredima osnovne škole nikako ne pridonosi bržem i učinkovitijem ovladavanju vještinom čitanja. Stalno suočavanje s novom nepoznatom riječi koja pritom nosi i novi fonološki plan pred dijete stavlja velike zahtjeve na razini njegove fonološke obrade. Izlaganjem djeteta riječima koje su već uskladištene u njegovome mentalnom leksikonu na temelju govorene komunikacije, i to u količini većoj od 12 puta u različitim pisanim kontekstima, ne pospješuje samo njegovu vještinu fonološkoga dekodiranja, već i olakšava prebacivanje na ortografsko dekodiranje, odnosno proces automatizacije čitanja. Time se osigurava brže povezivanje riječi s njezinim značenjem.

Čini se kako autori udžbenika namijenjenih učenicima prvih razreda, vjerojatno potaknuti teorijskom odrednicom o neraskidivosti rječnika i čitanja, usmjeravaju učenje čitanja na riječima koje su učeniku još nepoznate, za što podrazumijevaju i automatsko dohvaćanje njihova značenja i osiguravanje mjesta u mentalnome leksikonu. Nažalost, ovi procesi nisu nužno usporedni, odnosno istovremeni. Ovladavanje čitanjem iznimno

je složena fonološka vještina koja zahtijeva prethodno jezično znanje i sinkronizirano djelovanje niza drugih spoznajnih mehanizama zbog čega se ponekad usporava bogaćenje jezika novim elementima.

Kada se postave jasni obrazovni ciljevi usklađeni s djetetovim jezičnim mogućnostima, lakše će se postići potpuna usklađenost pisane građe namijenjene učenicima u prvim godinama obrazovanja s načelima *jasnoga jezika*. *Jasan jezik* će djetetu u razdoblju početnoga čitanja osigurati da lako pronađe što mu treba, razumije što je pronašao i uspješno upotrijebi novostečeno znanje.

Napomena

Ovaj je rad nastao u sklopu projekta *Razvoj inovativnog dijagnostičkog instrumentarija za rano prepoznavanje djece s disleksijom (KK.01.2.1.02.0167),* u sklopu poziva IRI 2 Istraživanje i razvoj: Povećanje razvoja novih proizvoda i usluga koji proizlaze iz aktivnosti istraživanja i razvoja - faza II.