https://doi.org/10.53656/for2024-01-03

DEVELOPING MATERIALS FOR ASSESSING READING LITERACY AND COMPREHENSION OF EARLY GRADERS IN BULGARIA AND ITALY

Prof. Vito Pirelli
Institute for Computational Linguistics "A. Zampolli"
Prof. Svetla Koeva
Bulgarian Academy of Sciences

Abstract. The paper presents the bilateral project *Assessing reading literacy and comprehension of early graders in Bulgaria and Italy*, its aims and expected results, as well as the principles underlying the creation of texts assessing reading literacy in Italian and Bulgarian. The ultimate goal of the project is to improve the literacy skills of primary school children through education. In order to contribute to the achievement of this goal, a thorough investigation focused on the assessment of reading literacy and comprehension among early school children in Bulgaria and Italy. The article focuses on the principles for preparing materials in Italian and Bulgarian for assessing reading literacy. The linguistic features used to predict reading ability are divided into four main groups: raw text, lexical, morpho-syntactic and syntactic features.

Keywords: reading literacy; assessing reading literacy; language technologies

Introduction

Educational systems across Europe strive to teach children core reading competencies. Literacy research and innovation has been at the forefront of this effort, supporting evidence-based practices for reading and language classes of schools everywhere. Good text reading and text understanding skills are key competences and an essential prerequisite to the uptake of high quality education. Reading fluency can predict all school marks in literacy-based subjects, with reading speed being the most important predictor. In the long term, students with early reading difficulties face serious problems with domain-general learning, school performance and social integration.

Ideally, education should be supported by continual observation of actual reading behaviour. Major international organisations such as Unicef (Chzhen et al. 2018) and OECD (OECD 2018) have lamented a regrettable shortage of large scale reading data. This explains why, in spite of substantial progress in our understanding of the

cognitive underpinnings of reading and the factors that occasionally make reading difficult to a child, there is still a long way to go before advancements in reading research can effectively be used in everyday educational practices.

In this paper we are going to present the bilateral (Bulgarian-Italian) project *Assessing reading literacy and comprehension of early graders in Bulgaria and Italy*, its goals and expected results, as well as the principles underpinning the creation of texts for assessing reading literacy in Italian and Bulgarian.

Project *Assessing reading literacy and comprehension of early graders in Bulgaria and Italy* in brief

The ultimate goal of the project is to increase the level of literacy and reading abilities of primary school children through education. To contribute to the achievement of this goal, we intend to carry out a comprehensive data-driven investigation focused on assessing reading literacy and comprehension in children of early school age in Bulgaria and Italy. In particular the project will:

– implement, and test assessment strategies for monitoring and evaluating the reading skills and the level of word comprehension of Bulgarian and Italian early graders;

– collect reading and comprehension evidence from the two populations of children according to the same battery of tools and comparable, highly-controlled test language materials;

– compare collected data across children from different age and social groups and with different languages (Bulgarian and Italian) and model the results;

– document and make available evidence-based procedures, protocols, and tools for reading and word comprehension assessment;

Research relies on both innovative and traditional methods for literacy research, coming from experimental psychology, psycholinguistics, and computational linguistics, which will be adapted to the specific purposes and scenarios envisaged by the project.

An ICT platform with a tablet front-ent has recently been developed by the Comphys Lab at CNR ILC in Pisa to provide accurate, evidence-based assessment of reading skills in early grade children. The platform, named Readlet (Taxitari et al. 2021; Crepaldi et al. 2022), can automatically collect, preprocess and analyse time-aligned multimodal reading data that include: voice recording, finger sliding time, time spent to answer comprehension questions, and number of correct answers. During a reading session with the tablet, a user is shown a one-page text displayed on the tablet touchscreen, and is asked to point to each word in the text as she reads it aloud, sliding the index finger of her dominant hand across the screen. The tablet keeps track of the sliding movements of the finger on the screen, while recording the reader's speech stream through the tablet built-in microphone. Both acoustic and haptic data are continuous in time,

while recorded finger trajectories are also continuous in space: i.e. they tend to cover text letters, punctuation marks and even blanks evenly, with a limited number of orthographic units being skipped. Data recorded by the tablet are first locally stored, then sent to a centralised server through an internet connection, where they are encrypted and anonymized for privacy protection, post-processed and time-aligned with the text.

A set of connected reading texts, containing short stories in Bulgarian and Italian, rigorously balanced and comparable in terms of their levels of readability and linguistic complexity were designed. During the project lifetime, texts will be administered to a population of young readers (from second to fifth grade level) through the ReadLet tablet, for aloud and silent reading data to be collected in both countries according to a controlled protocol and a shared methodology.

Using state-of-art NLP tools (Dell'Orletta et al. 2011; Koeva et al. 2020), the reading texts are annotated at different levels of linguistic analysis: from articulatory complexity (i.e. length and variety of consonant chunks) and phonological transparency, to part-of-speech tagging, lexical typicality (in terms of density and entropy of a word's lexical neighbourhood), orthotactic probability (as a function of a word's bigram and trigram probabilities), morphological complexity, token and type frequency, token's position and syntactic role in a sentence.

By aligning finger tracking data with the annotated reading text, we will be in a position to relate children's reading performance to specific linguistic factors across different grade levels, thereby gaining a better understanding of i) the basic mechanisms behind children's reading strategy, ii) what text factors make reading more difficult, iii) and ways to enhance a reader's strengths and overcome her weaknesses.

The interaction between efficient reading comprehension and word knowledge will be assessed by regressing reading scores on word comprehension scores as independent variables. As reading is a multidimensional process, involving a variety of cognitive, motor and even social factors, this type of analysis will allow us to control for all the covariates that may affect reading decoding and comprehension, and evaluate the contribution of each predictor as well as their interactions in terms of the goodness of fit of our regression models. In addition, data collection from different age groups will create the conditions for a fine-grained assessment of developmental profiles in the two languages.

Since both languages are purported to have fairly transparent orthographies, our analysis will include more subtle factors affecting grapheme-to-phoneme conversion such as the complexity and distribution of consonant chunks, the phonological systematicity of the rules mapping letter strings into sounds in the two languages, the density of their orthographic and phonological neighbourhoods and the comparative inflectional complexity for two major lexical categories such as verbs and nouns.

Developing reading texts for Italian and Bulgarian

Originally, five texts were first created for Italian (Taxitari et al. 2021). The texts are not parts or adapted versions of literary works; they were created from scratch for the specific research purposes of the project. The texts have a specific structure: each of them consists of five episodes, with the level of difficulty of the text increasing with each subsequent episode. After each episode, two questions are asked about the text, and there are four answer options from which you can choose the correct one. The content of the texts is tailored to the age for which they are intended. For example, one of the stories is about an alien boy who arrives on Earth and becomes friends with an earthly boy.

The texts in Italian were created according to a data-driven methodology for assessing levels of reading difficulty (Dell'Orletta et al. 2011). The approach is based on the assumption that text readability can be handled as a classification problem, specifically a binary classification intended to tell textual items that are easy to read from those that are difficult to read (Dell'Orletta et al. 2011, p. 75). The linguistic features that are used to predict readability are divided into four primary groups: raw text, lexical, morpho-syntactic, and syntactic features. Accordingly, the following stages of linguistic analysis are automatically carried out on the text being evaluated: tokenization, lemmatization, PoS tagging, and dependency parsing (Dell'Orletta et al. 2011, p. 76).

Raw text features include **Sentence length**, calculated as the average number of words per sentence, and **Word length**, calculated as the average number of characters per words.

Lexical features refer to the internal composition of the vocabulary of the text. For Italian, two different features were determined by comparing the results to a reference resource that included basic Italian word entries (De Mauro & Chiari 2016): a) the percentage of all unique words (types) in the text that are also present in this reference list (calculated on a per–lemma basis); b) the internal distribution of the occurring basic Italian vocabulary words into the usage classification classes of 'fundamental words' (very frequent words), 'high usage words' (frequent words) and 'high availability words' (relatively lower frequency words referring to everyday objects or actions and thus well known to speakers).

Morpho–syntactic features include calculation of **lexical density**: this feature refers to the ratio of content words (verbs, nouns, adjectives and adverbs) to the total number of lexical tokens in a text (Dell'Orletta et al. 2011, p. 77).

Syntactic features are several (Dell'Orletta et al. 2011, p. 77), not all are mentioned here:

– **Parse tree depth features** may indicate a rise in phrase complexity (Gibson 1998). The following measurements are directed to capture various facets of the parse tree depth: a) the depth of the entire parse tree, as determined by finding the longest path from the dependency tree's root to a leaf; b) the

average depth of embedded complement 'chains' governed by a nominal head and including either prepositional complements or nominal and adjectival modifiers; c) the probability distribution of embedded complement 'chains' by depth.

– **Relative ordering of subordinates with respect to the main clause**: sentences containing subordinate clauses in post-verbal rather than in pre-verbal position are easier to read (Miller and Weinert 1998).

– **Length of dependency**: the syntactic complexity of sentences can be predicted with measures based on the length of dependency links (Gibson 1998).

The Bulgarian texts are translations of the Italian ones. In their creation, the same principles were followed, but based on the translation of the Italian texts rather than on measurements that attest to the simplicity of the texts.The number of sentences is kept identical, as well as their type: a simple sentence is translated with a simple sentence, a complex sentence containing a complement clause with a complex sentence containing a complement clause, and so on and so forth. This way we avoided the splitting or joining of sentences that is a common practice in text translation.

Sentence length and **word length** can be calculated and compared with those in Italian texts based on the **total number of words and characters** in the texts. Table 1 presents a comparison of the number of characters and words in the first five sentences of the first episode of one of the texts. The comparison shows that the relatively low number of words in a sentence is preserved in Bulgarian, as is the tendency to use relatively short words, which in general mean words with a relatively simple morphological structure. The slightly greater average length of words in Bulgarian can be explained by the morphological structure of nouns, adjectives and some pronouns, in which the definite article is part of the word. Likewise, the relatively smaller average number of words in Bulgarian can be generally explained by the determiner in Italian, which is a separate word.

Table 1. Illustration of row text features of Bulgarian and Italian texts

Sentences	**Number of words**	**Number of characters**	**Average number of characters per word**
BG: Зиги е извънземно същество от далечна планета.	7	47	6.7
IT: Zigghi è un alieno da un pianeta lontano.	8	41	5.1
BG: Кръгъл е като футболна топка.	5	30	6
IT: È tondo come un pallone da calcio.	7	36	5.1

BG: Има две големи зелени очи и синьо-жълта козина.	8	47	5.9
IT: Ha due grandi occhi verdi e il pelo è blu e giallo.	12	51	4.3
BG: Зиги много обича да лети с червения си космически кораб.	11	60	5.4
IT: Gli piace molto volare con la sua nave spaziale rossa.	10	54	5.4
BG: Зиги иска да посети всички планети.	6	35	5.8
IT: Zigghi vuole visitare tutti i pianeti.	6	38	6.3
Average number of words per sentence (Bulgarian)	7.4	Average number of words per sentence in Italian	8.6
Average number of characters per words in the text (Bulgarian)	5.9	Average number of characters per words in the text (Italian)	5.2

To replicate the methodology followed for the Italian texts, **lexical features** for Bulgarian texts can be calculated in two ways: a) by comparing the vocabulary used in the translated texts with the **general Bulgarian lexis** (Koeva and Doychev 2022); and b) by calculating the **type/token ratio**: the ratio between the number of lexical types (different lemmas) and the number of tokens. For example, the type/token ratio for the first episode of one of the texts in Bulgarian is 0.67 (80 unique words and 119 tokens), while for Italian it is 0.61 (87 unique words and 141 tokens). Here, again, the general grammatical structures of the two languages have an influence on the results as well as the principles for tokenization and lemmatization. Nevertheless, the results are comparable. As for the usage of general lexis, the methodology includes a comparison of the vocabulary in textbooks at the particular stage of education as well as in a dictionary designed for this age group. Some Italian names have been replaced by names commonly used in Bulgaria, for example *Gianni* by *Ivo* and *Viola* by *Violetta*.

With respect to the morpho-syntactic features of the lexical density (the ratio of content words to the total number of tokens in a text), the lexical density of the same translated text in Bulgarian is 0.57 (69 unique content words and 119 tokens), while for the original Italian text it is 0.52 (74 unique content words and 141 tokens). Again, the results are pretty similar, and the differences might, in general, be explained by the different morphological structures of the determiners in the two languages.

Syntactic dependencies in Bulgarian and Italian can be calculated and visualised (as shown in Figures 1 and 2). The accepted principles for the simplicity of the texts were introduced during the creation of the Italian texts, and as far as possible, the

simplicity of the syntactic structure was followed in Bulgarian. Although the two languages, Bulgarian and Italian, share many morphological and syntactic features, they also have significant differences, which are manifested in their grammatical structure. The translation, for its part, represents a complex effort to convey the meaning as faithfully as possible while preserving the corresponding grammatical means of expression: active voice sentence with active voice sentence, imperative sentence with imperative sentence, etc.

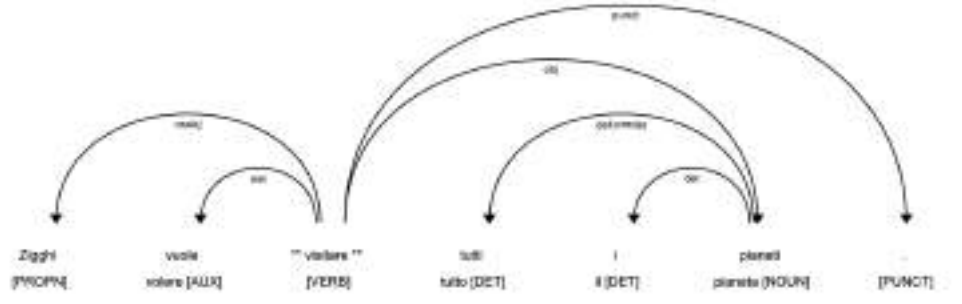

Figure 1. Dependency tree of the Italian sentence:
Zigghi vuole visitare tutti i pianeti.

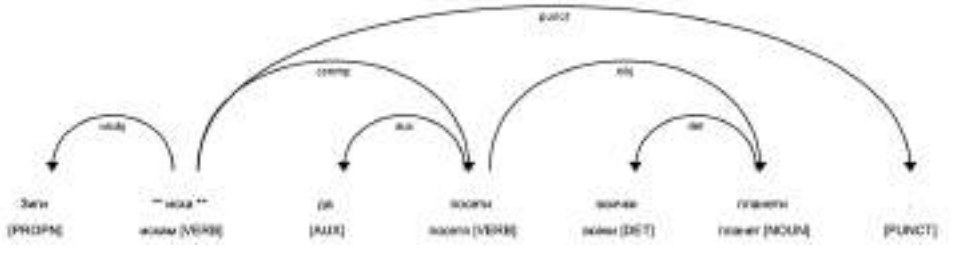

Figure 2. Dependency tree of the Bulgarian sentence:
Зиги иска да посети всички планети.

The comparison between the two dependency structures of the Bulgarian and Italian sentences shows that they are pretty similar: a matrix clause with the modal verb *want* is followed by a dependent clause with the verb *visit* heading the direct complement *all planets*. The two differences are as follows:

– *Volere* (want) is an Italian modal verb and can be followed directly by another verb, while the Bulgarian *verb искам* (want) requires a clause introduced by the conjunctive particle *да* (to).

– The noun phrase *all planets* in Italian contains the determiner *i: tutti i pianeti,* while in Bulgarian – not: *всички планети*.

The content of the texts has been chosen to attract and hold the children's attention. The original texts are not loaded with words denoting places, events or objects that are specific of the Italian culture and history, but they appear to be suitable for children of the same age.

Correspondences in morphological and syntactic representations are retained to a large extent, insofar as the grammatical structure of the two languages allows.

The correspondences at the lexical level are almost complete. In rare exceptions, it is necessary to replace a phrase in Italian with a complex word in Bulgarian (e.g. *blu e giallo: blue and yellow*, and *синьо-жълта:* blue-yellow), again due to the grammatical structure of the two languages.

To sum up, the translations of the Italian texts cannot be characterised as source-oriented or target-oriented translations, since we did not strive to preserve or introduce the cultural features of either language. This is because the source texts in Italian are structured in such a way that no background knowledge is required to understand their content.

Conclusion

By linking primary reading education with reading research and digital technology, we intend to bridge the current gap between school teachers' professional, subjective evaluation of early graders' reading skills and the quantitative models of reading proposed by research labs.

The importance of a project lies in several lines: first, the results of the project will outline the abilities of students from primary schools in Bulgaria and Italy to read fluently as well as to understand the meaning of the text they read. Secondly, thanks to the experiments done, prescriptions will be made for the lexical structure and the grammatical composition of the texts suitable for a certain age.

Acknowledgments. The research is developed under the project *Assessing reading literacy and comprehension of early graders in Bulgaria and Italy* (2023 – 2025), supported by the National Research Council of Italy and Bulgarian Academy of Sciences.

REFERENCES

CREPALDI, D., FERRO, M., MARZI, C., NADALINI, A., PIRRELLI, V., & TAXITARI, L. 2022. Finger movements and eye movements during adults' silent and oral reading. *Developing Language and Literacy*, pp. 443 – 471.

DE MAURO, T., & CHIARI, I., 2016. Il Nuovo vocabolario di base della lingua italiana. Internazionale. https://www. internazionale. it/opinione/tullio-de-mauro/2016/12/23/il-nuovo-vocabolario-di-base-della-lingua-italiana.

DELL'ORLETTA, F., MONTEMAGNI, S., & VENTURI, G., 2011. READ-IT: Assessing readability of Italian texts with a view to text simplification. *Proceedings of the second workshop on speech and language processing for assistive technologies*, pp. 73 – 83.

GIBSON, E. 1998. *Linguistic complexity: Locality of syntactic dependencies*. *Cognition*, vol. 68, no. 1, pp. 1 – 76.

KOEVA, S., N. OBRESHKOV, M. YALAMOV. 2020. Natural Language Processing Pipeline to Annotate Bulgarian Legislative Documents. *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 6988 – 6994, Marseille, France. European Language Resources Association.

KOEVA, S. & E. DOYCHEV. 2022. Ontology Supported Frame Classification. *Proceedings of the 5th International Conference on Computational Linguistics in Bulgaria (CLIB 2022)*, pp. 203 – 213, Sofia, Bulgaria. Department of Computational Linguistics, IBL – BAS.

MILLER, J. & R. WEINERT, 1998. *Spontaneous spoken language. Syntax and discourse*. Oxford, Clarendon Press.

TAXITARI, L., CAPPA, C., FERRO, M., MARZI, C., NADALINI, A., & PIRRELLI, V., 2021. Using mobile technology for reading assessment. *2020 6th IEEE Congress on Information Science and Technology (CiSt)*, pp. 302 – 307.

✉ **Prof. Dr. Vito Pirelli**
ORCID iD: 0000-0002-5581-7451
Institute for Computational Linguistics "A. Zampolli"
1, Via Giuseppe Moruzzi
56124 Pisa, Italy
E-mail: vito.pirrelli@ilc.cnr.it

✉ **Prof. Dr. Svetla Koeva**
ORCID iD: 0000-0001-5947-8736
Department of Computational Linguistics
Institute for Bulgarian Language "Prof. Lyubomir Andreychin"
Bulgarian Academy of Sciences
52, Shipchenski prohod Blvd., Bldg 17
Sofia, Bulgaria
E-mail: svetla@dcl.bas.bg