

ReaderBench: A Multi-lingual Framework for Analyzing Text Complexity

Mihai Dascalu^{1,2}✉, Gabriel Gutu¹, Stefan Ruseti¹, Ionut Cristian Paraschiv¹,
Philippe Dessus³, Danielle S. McNamara⁴, Scott A. Crossley⁵,
and Stefan Trausan-Matu^{1,2}

¹ University Politehnica of Bucharest, Splaiul Independenței 313, 60042 Bucharest, Romania
{mihai.dascalu, gabriel.gutu, stefan.ruseti, ionut.paraschiv, stefan.trausan}@cs.pub.ro

² Academy of Romanian Scientists, Splaiul Independenței 54, 050094 Bucharest, Romania

³ Laboratoire des Sciences de l'Éducation, Univ. Grenoble Alpes, F38000 Grenoble, France
philippe.dessus@univ-grenoble-alpes.fr

⁴ Institute for the Science of Teaching and Learning, Arizona State University, Tempe, USA
dsmcnama@asu.edu

⁵ Department of Applied Linguistics/ESL, Georgia State University, Atlanta 30303, USA
scrossley@gsu.edu

Abstract. Assessing textual complexity is a difficult, but important endeavor, especially for adapting learning materials to students' and readers' levels of understanding. With the continuous growth of information technologies spanning through various research fields, automated assessment tools have become reliable solutions to automatically assessing textual complexity. *ReaderBench* is a text processing framework relying on advanced Natural Language Processing techniques that encompass a wide range of text analysis modules available in a variety of languages, including English, French, Romanian, and Dutch. To our knowledge, *ReaderBench* is the only open-source multilingual textual analysis solution that provides unified access to more than 200 textual complexity indices including: surface, syntactic, morphological, semantic, and discourse specific factors, alongside cohesion metrics derived from specific lexicalized ontologies and semantic models.

Keywords: Multi-lingual text analysis · Textual complexity · Comprehension prediction · Natural Language Processing · Textual cohesion · Writing style

1 Introduction

Two important and cumbersome tasks, which often face many teachers, are selecting reading materials suitable for their students' levels of understanding, and assessing their written productions (e.g., essays, summaries). In order to support both tasks, *ReaderBench* [1], a multilingual, open-source framework centered on discourse analysis, was developed. From an architectural perspective, as shown in Fig. 1, our framework comprises three layers: (a) *linguistic resources* that provide solid language background knowledge and can be used to train the semantic models and compute various measures;

(b) *linguistic services* used to process and append semantic meta-information to text resources, and (c) *linguistic applications* that rely on machine learning and data mining techniques, and are designed for various educational experiments and visualizations. *ReaderBench* implements various metrics and categories of textual complexity indices that can be used to leverage the automated classification of datasets in multiple languages, such as English [2], French [3], Romanian [4] and Dutch [5].

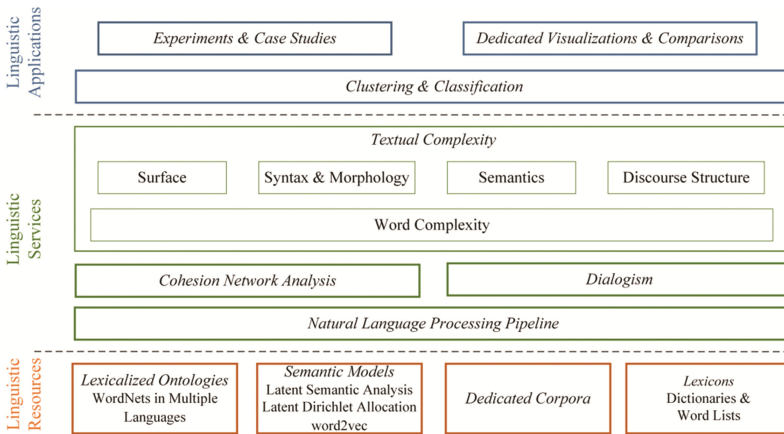


Fig. 1. *ReaderBench* processing architecture.

2 Description of Textual Complexity Indices

More than 200 textual complexity indices computed by the *ReaderBench* platform have been used in a number of experiments. *ReaderBench* integrates a multitude of indices, discussed briefly below, ranging from classic readability formulas, surface indices, morphology and syntax, as well as semantics and discourse structure.

Surface Indices. These are the simplest measures that consider only the form of the text. This category includes indices such as sentence length, word length, the number of unique words used, and word entropy. All these indices rely on the assumption that more complex texts contain more information and, inherently, more diverse concepts.

Word Complexity Indices. This category of indices focuses on the complexity of words, but goes way beyond their form. Thus, the complexity of a word is estimated by the number of syllables and how different the flecional form is from its lemma or stem, considering that adding suffixes and prefixes increases the difficulty of using a given word. Moreover, a word’s complexity is measured by considering the number of potential meanings derived from the word’s senses available in WordNet, as well as a word’s specificity reflected in its depth within the lexicalized ontology.

Syntactic and Morphologic Indices. These indices are computed at the sentence level. The words' corresponding parts of speech and the types of dependencies that appear in each sentence can be used as relevant measures, reflective of a text's complexity. In addition, named entity-based features are tightly correlated with the amount of cognitive resources required to understand the given text.

Semantic Cohesion Indices. Cohesion plays an important role in text comprehension and our framework makes extensive usage of Cohesion Network Analysis. *ReaderBench* estimates both local and global cohesion by considering lexical chains, different semantic models (semantic distances in different multilingual WordNets, LSA – Latent Semantic Analysis, LDA – Latent Dirichlet Allocation, and Word2Vec), as well as co-reference chains.

Discourse Structure Indices. Specific discourse connectives and metrics derived from polyphonic model of discourse [1], which considers the evolution of expressed points of view, provide additional valuable insights in terms of the text's degree of elaboration. Word features and vectors from the integrated linguistic resources are also used to reflect specific discourse traits.

3 Validation Experiments

Multiple experiments have been performed to validate *ReaderBench* as a multi-lingual text analysis software framework. This section focuses on the latest and most representative experiments conducted in English, French, Romanian, and Dutch languages. The *first experiment* [2] was performed on a set of 108 argumentative essays written in English and timed to 25 min. For the analysis, only essays that contained three or more paragraphs were considered in order to use global cohesion measures reflective of inter-paragraph relations. Individual difference measures such as vocabulary knowledge and reading comprehension scores were assessed. The results showed that writers with stronger vocabulary knowledge used longer words with multiple senses and higher entropy, but also created more cohesive essays. Also, students with higher reading comprehension scores created more cohesive and more lexically sophisticated essays, using longer words, and with higher entropy.

The *second experiment* [3] relied on a set of 200 documents collected from primary school French manuals. The documents were pre-classified into five complexity classes mapped onto the first five primary grade levels of the French national education system. A Support Vector Machine (SVM) was used to classify the documents. The pre-trained model was used to determine the complexity for an additional set of 16 documents that were manually classified into three primary grades. Students belonging to the three classes had to read the texts and answer a posttest. Correlations between the textual complexity factors' scores and the students' average scores were computed. This allowed the computation of the impact for each factor in calculating the reliability of prediction of the textual complexity score for a given document.

The *third experiment* [4] was conducted on a set of 137 documents written in Romanian language. The documents were collected from two time periods, 1941–1991 and

1992-present, and two regions, Bessarabia and Romania. The first period altered the Romanian language spoken in the country because of the implementation of the Russian language into the education system of Bessarabia. The aim of the experiment was to determine whether differences between the two regions and the two time periods could be observed in relation to the complexity of written texts. The analysis showed that more elaborated texts were created in the second period for both Bessarabia and Romania, while more unique words have been used in the second period for Bessarabia, but remained the same for Romania. The semantic cohesion of the texts increased over time, but no significant differences were observed between the two regions.

The *fourth experiment* [5] was run on a set of 173 technical reports written in Dutch language belonging to high or low performance students. Due to the length of the documents, a multi-level hierarchical structure was automatically generated based on the section headings. The experiment showed that students who received higher scores had longer reports, but also greater word entropy. They used more pronouns, discourse connectors and unique words, but also had lower inner cohesion scores per paragraph which is indicative of more sophisticated paragraphs.

4 Conclusion

Many pedagogical scenarios can fully integrate the use of *ReaderBench*, thanks to its versatility. The wide range of textual assessment features can support both teachers' assessment and learners' writing self-regulation. Moreover, multiple learning contexts take advantage from *ReaderBench's* support: either individual textual production and reflection, or collaborative knowledge building.

The presented experiments support the *ReaderBench* framework for determining the textual complexity of texts written in English, French, Romanian, and Dutch languages. Other languages, such as Spanish, Italian, and Latin are also partially supported. To our knowledge, *ReaderBench* is a unique *multilingual* system that provides access to a wide range of textual complexity indices and to various textual cohesion analyses.

Acknowledgments. This research was partially supported by the FP7 2008-212578 LTFL project, by the 644187 EC H2020 RAGE project, by the ANR-10-blanc-1907-01 DEVCOMP project, as well as by University Politehnica of Bucharest through the "Excellence Research Grants" Program UPB-GEX 12/26.09.2016.

References

1. Dascalu, M.: Analyzing Discourse and Text Complexity for Learning and Collaborating. SCI, vol. 534. Springer, Cham (2014)
2. Allen, L.K., Dascalu, M., McNamara, D.S., Crossley, S., Trausan-Matu, S.: Modeling individual differences among writers using ReaderBench. In: EduLearn 2016, Barcelona, Spain, pp. 5269–5279. IATED (2016)

3. Dascalu, M., Stavarache, L.L., Trausan-Matu, S., Dessus, P., Bianco, M.: Reflecting comprehension through French textual complexity factors. In: 26th International Conference on Tools with Artificial Intelligence, ICTAI 2014, Limassol, Cyprus, pp. 615–619. IEEE (2014)
4. Gifu, D., Dascalu, M., Trausan-Matu, S., Allen, L.K.: Time evolution of writing styles in Romanian language. In: ICTAI 2016, San Jose, CA, pp. 1048–1054. IEEE (2016)
5. Dascalu, M., Westera, W., Ruseti, S., Trausan-Matu, S., Kurvers, H.: ReaderBench learns Dutch: building a comprehensive automated essay scoring system for Dutch. In: André, E., Baker, R., Hu, X., Rodrigo, M., du Boulay, B. (eds.) AIED 2017. LNCS, vol. 10331, pp. 52–63. Springer, Cham (2017)