RUPKATHA JOURNAL
ON INTERDISCIPLINARY STUDIES IN HUMANITIES
ISSN 0975-2935

A Model Text Recommendation System for Engaging English Language Learners: Facilitating Selections on CEFR

INTERACTIVE ARTICLE COVER

About the Journal

Journal DOI	https://dx.doi.org/10.21659/rupkatha
Journal Home	www.rupkatha.com
Indexed by	Scopus Web of Science: Emerging Sources Citation Index (ESCI) DOAJ
Journal Metrics	CiteScore 2020: 0.2 | SJR 2020: 0.162 | SNIP 2020: 0.193 | JCI 2020: 0.50

About the Issue

Issue	**Volume 14, Number 3, 2022**
Editor	Tirtha Prasad Mukhopadhyay
Affiliation	Universidad de Guanajuato
Issue DOI	https://doi.org/10.21659/rupkatha.v14n3
TOC	https://rupkatha.com/v14n3.php

About the Article

Title	**A Model Text Recommendation System for Engaging English Language Learners: Facilitating Selections on CEFR**
Author/s	**Adelina Escobar-Acevedo[1], Josefina Guerrero-García[1], Rafael Guzmán-Cabrera[2]**
Affiliation	[1]Benemérita Universidad Autónoma de Puebla, México. [2]Universidad de Guanajuato, Mexico.
Funding	**No funding received. Published free of any charge.**
Article DOI	**https://doi.org/10.21659/rupkatha.v14n3.17** **Pages:** 1-8
Abstract	https://rupkatha.com/v14n317
Full-text PDF	https://rupkatha.com/V14/n3/v14n317.pdf
Article History	**First Published: 17 October 2022**
Article Impact	**Check Dynamic Impact**
Copyright	Aesthetics Media Services
Licensing	Creative Commons Attribution Non-Commercial 4.0

A Model Text Recommendation System for Engaging English Language Learners: Facilitating Selections on CEFR

Adelina Escobar-Acevedo[1], Josefina Guerrero-García[1], Rafael Guzmán-Cabrera[2]
[1]Benemérita Universidad Autónoma de Puebla, Facultad de Ciencias de la Computación, México. Email: adelina.escobar@alumno.buap.mx
[2]Universidad de Guanajuato, Campus Irapuato-Salamanca, División de Ingenierías, Departamento de Ingeniería Eléctrica.

Abstract

A pedagogically informed multimodal education system is defined by how well reading tasks are assigned to students in a contemporary classroom. A source that becomes a provider of readings is the web, where it is possible to find information on practically all areas of knowledge and in a wide variety of languages. However, selecting the appropriate material for the level and theme becomes a tedious job to which language teachers must devote a significant amount of their time. Selecting suitable readings to accompany the teaching-learning process is thus not a 'trivial' task. Basic-level texts for language competence are easy to recognize and obtain but as is seen in case of the Common European Framework of Reference for Languages recommendations (CEFR), selection of appropriate texts that impart language competencies, especially of vocabulary and grammar at higher levels of communicativeness, selection becomes increasingly complex for teachers. Furthermore, the suggested readings should be raked by complexity in accordance with student capabilities. We suggest, that automatic classifiers based on CEFR levels may help in this process of selections from the already available corpora of authentic texts on the web. The existing facility of access of readers to such material on the web may come to the aid of automated classifiers. Teachers use interest to motivate reading in classrooms, but automatic recommendation systems will allow specific or even individualized recommendations. The authors explore the impact of such multimodal methods on the acquisition of better linguistic and communicative skills.

Keywords: English Language Learners, CEFR Language level, Linguistic Features, Text Complexity.

Educational Strategies in English [L2] Teaching

Around the world, bilingualism is a desirable skill. Linguistic globalization is driven primarily by the English language because of historical and socioeconomic factors such as make learning English, or English L2 acquisition, an important goal for the education sector in almost every part of the world where non-English speaking people live and engage in global trade and economic processes. The English language plays a fundamental role in general communication and the exchange of knowledge. It has emerged as a common language for science, literature, music, business, diplomacy, and migration (Hernández-Fernandez y Rojas, 2018) in the context of the recent development of geopolitical sectors like the European Union, the Middle East or the Far East economic blocks. Acquisition of English as a second language is crucial throughout Latin America and Africa and has managed to arouse the interest of governments in establishing public policies to implement all kinds of pedagogical techniques for its entire learning population. At a macro level, there is a demonstrated correlation between the skills of using English for a population and the economic performance of the same. At the micro (individual) level, English is a recognized tool for economic and academic development (Grin, 2014), increasing a person's income by upto 25% in developing countries (McCormick, 2013). Reading, a fundamental activity in language learning, is encouraged and practiced daily in every classroom. Reading is a very complex process involving a wide variety of abilities and strategies expected for each text. As Birch (2014) indicates, "Lowest level processing strategies are as crucial to reading success as the higher-level cognitive strategies."

Reading curricula are based on students' needs, as text interaction involves tags, instructions, newspapers, and academic articles. However, selecting suitable readings for levels of complexity appropriate for an accompanying learning process is not a trivial activity. As language acquisition is a progressive process, texts for a basic level should be easy to read and understand; as the level increases, the complexity of the text should also increase commensurately. For example, the structure of any language with an extensive vocabulary and encompassing as it does, more specialized topics and more formal language at superior or complex levels of communicative organization, would present variables of greater organization and complexity in more upper-level texts which deal with complex shades and tones of information and significance. Language levels are aligned to a reference framework which clarifies students' competencies in each level. One such popular reference framework that is designed for variable discourse competence is, of course, the widely used CEFR, which indicates six levels: basic user (A1, A2), independent user (B1, B2), and proficient user (C1, C2). Each level usually takes up to twice the time invested in the predecessor level for the acquisition of competencies (Council of Europe, 2001); neither are these levels of competence acquisition linear.

The Zone of Proximal Development for English Language Learners

A fundamental concept in social constructivism is *scaffolding*, linked to the learner's Zone of Proximal Development (ZPD) indicated by Vygotsky (1978). There is a range of tasks and activities which the student can achieve through support, but which may lie beyond that student's current abilities (Yang and Wilson, 2006), and therefore requires learning, with all kinds of active intelligence modules available in the environment. The support provided by others (teachers, peers, tools) enables students to increase their performance; students benefit in crossing over from the liminal threshold of the known to the unknown, where they can expand their knowledge gradually, and without getting overwhelmed by the increasing amount of information compiled at a superior level. This kind of theoretical trajectory explains the process of intellectual maturation that is completed at a certain threshold but also those that continue to mature and develop in a constructivist way for the individual. Unlike traditional conceptions of development, which are measured by retrospective evaluation, Vygotsky also provides the possibility of a prospective analysis. The ZPD allows projecting the learner's immediate future and dynamically evolving state of learning, identifying what the learner interprets by way of development and also what is predictable to an extent, during his or her evolution through a future retinue. Some additional paradigmatic inferences can also be made: for example learning a second language facilitates mastery of higher forms of the native language (Torga, sf), Vygotsky affirms this idea when he postulates that the child learns to consider his/her language as a particular system among many other existing systems, and to view its language phenomena under more general categories, and this leads to an understanding of human linguistic operations in a kind of reflective, metacognitive way (Vygotsky 1995; 149-150). Although leading students to a threshold is desirable, overpassing it may provoke the opposite effect, demotivating learners; this is the reason why ranking text levels correctly is of fundamental importance from the psychological point of view of the learner. An experiment designed by Laufer (1989) indicated that students should know around 95% of the text's vocabulary for reading comprehension. It is furthermore known that keeping students' attention aroused and alive is one of the other difficulties for threshold learners.

Motivation as an Engine for Learning

It is impossible to separate the learning process from affective mechanisms (Pellaud et al., 2021). Interest has been extensively studied as a condition for learning. In the field of education, research focuses on two types of interest: situational and individual interest. The first is caused by the environment and can positively influence reading comprehension, contributing to essential aspects such as inference, information integration, and learning improvement. The second is related to predisposition and is helpful for attention, recognition, and other desired effects. Teachers can influence both (Hidi et al., 2006). Reading is an activity with a high cognitive load. Cognitive load implies that for "the amount of mental work involved in a task – the more work, the more reluctant the reader is to do it" (Grabe, 2009). Thus, situational interest is one of the resources used by teachers of English as a Foreign Language (EFL) in the classroom, mainly with children and adolescents. Hence, individual interest can be exploited to engage students to keep reading through an artificially generated recommendation system that provides text from the preferred topics.

Disadvantages of Material Selection for Classroom Scenarios

To define labels for the material for each level, in the context of labels already identified in CEFR levels, language specialists must read the text and the exercises attached. There are many features and assumptions to consider here as this tagging is related to language. Some texts are difficult to tag, especially if, due to their characteristics, they are on the threshold between two levels. There is disagreement in the literature regarding tagging through threshold-level texts and preferences. The difficulty of categorization is common among educators: what if a certain category of vocabulary is not yet easily adapted to the cognitive level of the students in a learning scenario? While recommending reading involving such textual uncertainties, we could revert back to the paradigm followed in conventional human teaching modules where texts are simply adapted manually to exclude complex vocabulary or grammar and colloquial expressions. This process is known as text simplification. Thus, creators protect corpora as they are costly in time, and effort, and rely on specialist human resources. An extra ranking action occurs when teachers actually present reading material to language-learning students. Teachers' efforts are focused on the average class situation. Language classrooms are a merge of students whose skills generally belong to an average level of competence reproduction but the actual level of these L2 students is not at all uniform. As Birch (2014) mentioned, factors such as incomplete knowledge of English, interference, and missing English processing strategies that the teacher must consider while working on English Language Learners (ELL). Considering the basis of the competence of every individual learner alone makes it possible to address the requirement of the vocabularies for a particular student thus making it possible for the literacy transfer to be of a positive, rather than neutral or even negative effect. In the entire program of providing adequate texts to language learners, one persistent problem therefore is the lack of standard-level-annotated corpora. Most studies are based on data collected from learning resources. For example, researchers use language books to extract readings, but they face many obvious restrictions; even some sources prohibit the electronic storage of their material. During the development of this work, the authors noticed an additional complication. Most of the CEFR material available is dedicated to children, with fewer tagged resources for adults. The material should be focused on the student level as well as the interests of children which are not always the same for again, an adult population of learners.

Related Work

Some online tools allow for evaluating text complexity and assigning a score. Web tools to evaluate text complexity are better for native texts. The main disadvantage of these kinds of tools is that they provide approaches to CEFR levels only, as they use vocabulary lists or readability formulas. Readability formulas are used broadly by writers, editors, and many other branches of content production. Readability formulas are widely used as they are domain-independent; in simple words, they are not built for specific topics, which also means they try to capture general information. Famous metrics include word length and word count, for example, but there are multiple of assumptions to consider here. There is blind trust in readability formulas to set forth a complexity label to text. In practice, the popular Flesch Kinkaid Grade Level (FKGL) is commonly used by language teachers to grade language-level readings or compare English textbooks Zhang (2016), Khodadady and Mehrazmay (2017), Cárcamo Morales (2020) and Gizatulina et al. (2020). Nevertheless, it is known that there are differences between text complexity for native speakers and second-language learners (Natova, 2019). Predictions on the difficulty of texts were studied by Nahatame, (2020) using eye movements and readability formulas. Including one readability formula targeted to Language Learners and included in a tool called Coh-Metrix (Graesser et al., 2017). Language is linked to communication media. Thankfully language technologies are developing quickly (Guzman Cabrera, 2022). Now, advances in language technology are also evolving with increasing growth of applications and innovations for a new kind of educational environment, especially after the pandemic alerted us to various constraints and opened possibilities of newer learning environments (and adaptations) of readability formulas for CEFR modules. For example, text analysis allows extracting enough information to perform automatic tasks like text classification. Nowadays, it is possible to predict students' language levels through their essays (Arnold et al., 2018) and tag new texts from Wikipedia (Wilkens et al., 2018) and even identify students' reading comprehension through their responses (McCarthy et al., 2020). More exposure time to internet makes such applications of machine learning aids more and more useful and amenable.

Some authors have tried to match school grade levels for native readers with language levels for second or foreign learners (Xia et al.,2016). These kinds of alignments also constitute an approach. Recent research indicates that superficial features used in common readability formulas are insufficient for CEFR text complexity classification. There is a need to add more linguistic features. One example is Kurdi's research (2020); he used CEFR-labeled text from multiple sources to create a classifier using linguistic features. For synthesis, there are resources from second language learners' productions, but those are suitable for the evaluation of students, not texts. There are native readings that are not ranked for second language learners even if there are some attempts to set a match, and so there is a lack of standard corpus for the specific task of CEFR classification. In text difficulty classification, readability formulas are also used for second-language learners. Recently, works such as Kurdi (2020) have added linguistic features; this work follows such recommendations.

Proposed recommendation system as an engagement tool to motivate reading

We pursue the creation of a recommendation system that succeeds in identifying suitable text levels for language learners and motivating them to read through user-preferred topics. Generally, the main disadvantage of the development of such recommendatory systems is the lack of free and extensive expert-tagged corpora at CEFR levels. In order to build an adequate computational model for language learning (EFL), it is mandatory to use a tagged corpus. In an application algorithm that we developed (Escobar-Acevedo and Guerrero-García, in press) the corpus consisted of British Council readings downloaded from the web pages of each level A2, B1, B2, and C1. British Council resources are public and accessible and provide credit[i]. Readings are related to everyday life, including academic and professional tips, and are based on such well-defined corpora predictions as a chosen conference or a specific round of studies. For our study, four of the five CEFR levels were used[ii]. Level A1 was not considered because of its format. On this level, documents are focused on information extraction and consists of schedules, presentation cards, and chat conversations. There are few or no sentences. One B2 document was excluded for the same reason, a chat conversation. From each document, considering our model, only text without questions or exercises was considered without preprocessing. Document length increases as the level does for obvious reasons of the increasing complexity of associations on superior levels of language use or competence.

The Coh-Metrix is a free tool to perform deep text analysis (Graesser et al., 2017). It provides 106 metrics, including computing simple numbers such as paragraph counts to complex linguistic features counts such as cohesion and readability metrics. Coh-Metrix analysis is divided into eleven categories according to its documentation. Each document in the corpus that we considered for the recommendational algorithm was analyzed with Coh-Metrix. Due to the straightforward interpretation, a decision tree was constructed. Decision trees use heuristic algorithms to determine decision attributes. They are dependent on the training set. Typically, each node of the tree takes a decision and finishes in a leaf, which indicates the tag that is to be assigned to objectify the level of its usability. For our model (Escobar-Acevedo and Guerrero-García, in press), we preferred using the J48 from Weka to build the decision tree. The predicted level is the tag assigned automatically to each text based on a decision tree. A test set was used to determine the exactitude of the model by using the original tag against those predicted. Of 106 attributes, seven were selected in our model, and five were related directly to words. It is worth mentioning that concrete words are meaningful and can be related to an image (Morett, 2019). As a text containing more concrete words is more easily processed, the incidence of concrete words in a text is related to the ease of reading it. Identifying connectors allows the reader to understand the purpose behind reading, which is also related to ease of reading.

Discussion and conclusions

Our model of artificial text recommendation allows us to build a simple classifier for texts. Corpus used for training is available for free, and it is already tagged on CEFR levels. Linguistic attributes are obtained from the free tool called Coh-Metrix that can be used via the web, as we mentioned above. The result is a very human-understandable model that overcomes randomness. It can be used by teachers, students, and the general public to rank texts at CEFR levels. Because the model was trained using linguistic attributes, the authors claim that this model is domain-independent, which means that it is not restricted to specific topics; but further experiments must be done. On the attributes selected, there were 106 indices; the algorithm did not select any of the three available readability metrics in Coh-Metrix for this task (FRE, FKGL, RDL2). Readability formulas are commonly used to rank text for native readers since those were created for that purpose. The result of our experiment reveals a range of difference in texts aimed for English learners. We could allow for ranking of texts on language levels automatically. That allows teachers and students to use text on an adequate level of language. Texts can be ranked by particulars or used massively in Massive Open Online Courses (MOOCs) or Intelligent Tutoring Intelligent Systems (STIs). Future work will be focused on gaining exactitude on two paths, first by creating new models with other algorithms and second by expanding results under new corpora. There is an opportunity of creating recommendation systems for materials suitable in educational environments for more differentiated learners' ages and languages.

What this means is that artificial selections may be used for language learning in the future in ways that are more agile and flexible, and which make tutorial intervention a less demanding task. The selection of texts for reading and engaging could be done with a certain degree of speed and accuracy. This would bring about a radical transformation of the manner in which learning retinues are applied in the academy and even outside of it. It could assist in activated learning scenarios by creating a computational tool for interested learners, and thus emerge as a great facilitator for communicative competencies. Its social effects could be further analysed in a different kind of assessment.

Declaration of Conflict of Interests

The author(s) declared no potential conflicts of interest.

Funding

No funding has been received for the publication of this article. It is published free of any charge.

Notes

---

[i] https://www.britishcouncil.org/terms?_ga=2.259359498.1464239874.1651466817-1576384399.1648489823

[ii] https://learnenglish.britishcouncil.org/skills/reading

References

Arnold, T., Ballier, N., Gaillat, T., and Lissón, P. (2018). *Predicting CEFRL levels in learner English on the basis of metrics and full texts*. arXiv preprint arXiv:1806.11099.

Birch, B. M. (2014). English L2 reading: Getting to the bottom, second edition. In *English L2 Reading: Getting to the Bottom, Second Edition*. https://doi.org/10.4324/9781410614933

Cárcamo Morales, B. (2020). Readability and types of questions in Chilean EFL high school textbooks. *TESOL Journal*, *11*(2), 1–15. https://doi.org/10.1002/tesj.498

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge University Press.

Escobar-Acevedo, A., and Guerrero-García, J. (in press). Hacia la construcción de un Sistema Tutor Inteligente: Estudio piloto de selección de atributos. In *Lenguaje, conocimiento y tecnología educativa: nuevos enfoques de aplicación*. Benemérita Universidad Autónoma de Puebla.

Gizatulina, D., Ismaeva, F., Solnyshkina, M., Martynova, E., and Yarmakeev, I. (2020). Fluctuations of text complexity: the case of Basic State Examination in English. *SHS Web of Conferences*, *88*.

Grabe, W. (2009). Reading in a second language: Moving from theory to practice. In *Studies in Second Language Acquisition* (Vol. 32, Number 4).

Graesser, A. C., McNamara, D. S., and Louwerse, M. M. (2017). *Coh-Metrix*. http://cohmetrix.com/

Grin, F. (2014). *50 years of economics in language policy Critical assessment and priorities*.

Guzman Cabrera, R. (2022). Sentiment Analysis for a Humanist Framework: How Emotions are Recognized and Interpreted in the Age of Social Media. *Rupkatha Journal*, *14*(2), 1–9. https://doi.org/https://dx.doi.org/10.21659/rupkatha Journal

Hernández-Fernandez, J., and Rojas, J. (2018). English Public Policies in Latin America: Looking for Innovation and Systemic Improvement in Quality English Language Teaching. In *English Public Policies in Latin America: Looking for Innovation and Systemic Improvement in Quality English Language Teaching*. British Council.

Hidi, S., Renninger, K. A., and Hidi, S. (2006). *The Four-Phase Model of Interest Development The Four-Phase Model of Interest Development*. *1520*(2006). https://doi.org/10.1207/s15326985ep4102_4

Khodadady, E., and Mehrazmay, R. (2017). Evaluating two high intermediate EFL and ESL textbooks:

a comparative study based on readability indices. *Sociology International Journal*, *1*(3). https://doi.org/10.15406/sij.2017.01.00016

Kurdi, M. Z. (2020). Text Complexity Classification Based on Linguistic Information: Application to Intelligent Tutoring of ESL. *Journal of Data Mining & Digital Humanities*, *2020*. http://arxiv.org/abs/2001.01863

Laufer, B. (1989). What percentage of text-lexis is essential for comprehension? *Special language: From humans thinking to thinking machines*, *316323*(January 1989).

McCarthy, K. S., Allen, L. K., and Hinze, S. R. (2020). Predicting Reading Comprehension from Constructed Responses: Explanatory Retrievals as Stealth Assessment. In I. I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, y E. Millán (Eds.), *International Conference on Artificial Intelligence in Education* (Vol. 12164, pp. 197–202). Springer International Publishing. https://doi.org/10.1007/978-3-030-52240-7_36

McCormick, C. (2013). *Countries with Better English Have Better Economies*. https://hbr.org/2013/11/countries-with-better-english-have-better-economies

Morett, L. M. (2019). The Power of an Image: Images, Not Glosses, Enhance Learning of Concrete L2 Words in Beginning Learners. *Journal of Psycholinguistic Research*, *48*(3), 643–664. https://doi.org/10.1007/s10936-018-9623-2

Nahatame, S. (2020). Text readability and comprehension processes during L2 reading: A computational and eye-tracking investigation. *Conference of the American Association for Applied Linguistics (AAAL)*.

Natova, I. (2019). Estimating CEFR reading comprehension text complexity. *The Language Learning Journal*, *0*(0), 1–12. https://doi.org/10.1080/09571736.2019.1665088

Pellaud, F., Shankland, R., Blandenier, G., Dubois, L., Gey, N., Massiot, P., and Gay, P. (2021). *The Competencies That School-Leavers Should Possess in Order to Meet the Challenges of the 21st Century*. *6*(May), 1–12. https://doi.org/10.3389/feduc.2021.660169

Torga, C,M. (sf) *Vygotsky y Krashen: "Zona de desarrollo próximo y el aprendizaje de una lengua extranjera"*. Universidad Nacional del Comahue.

Vygotsky, L. S. (1978). *Mind in society: The development of higher mental processes*. E. Rice, Ed & Trans.

Vygotsky, L.S. (1995). Pensamiento y Lenguaje. Ediciones Fausto.

Wilkens, R., Zilio, L., & Fairon, C. (2018). SW4ALL: a CEFR-Classified and Aligned Corpus for Language Learning. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 365–370. http://www.lrec-conf.org/proceedings/lrec2018/pdf/1012.pdf

Xia, M., Kochmar, E., & Briscoe, T. (2016). *Text Readability Assessment for Second Language Learners*. *January*, 12–22. https://doi.org/10.18653/v1/w16-0502

Yang, L., and Wilson, K. (2006). Second Language Classroom reading: A social constructivist Approach. *The reading Matrix: CALL Technologies and the Digital Learner*, *6*(3), 364–372.

Zhang, R. (2016). A Coh-Metrix Analysis of Two Textbooks: Successful English for Vocational Colleges and Vocational College English (An Integrated Skills Course). *US-China Foreign Language*, *14*(5), 351–356. https://doi.org/10.17265/1539-8080/2016.05.004