Journal of Language Horizons

Journal of Language Horizons, *Alzahra University*
Volume 6, Issue 1, Winter and Spring 2022 (Biannual – Serial No. 11)

Application of Computational Linguistics to Predicting Language Proficiency Level of Persian Learners' Textbooks

Research Article
pp. 29-52

Masood Ghayoomi[1]

Received: 2020/08/15 Accepted: 2021/02/09

Abstract
One subfield of assessment of language proficiency is predicting language proficiency level. This research aims at proposing a computational linguistic model to predict language proficiency level and to explore the general properties of the levels. To this end, a corpus is developed from Persian learners' textbooks and statistical and linguistic features are extracted from this text corpus to train three classifiers as learners. The performance of the models vary based on the learning algorithm and the feature set(s) used for training the models. For evaluating the models, four standard metrics, namely accuracy, precision, recall, and F-measure were used. Based on the results, the model created by the Random Forest classifier performed the best when statistical features extracted from raw text is used. The Support Vector Machine classifier performed the best by using linguistic features extracted from the automatically annotated corpus. The results determine that enriching the model and providing various kinds of information do not guarantee that a classifier (learner) performs the best. To discover the latent teaching methodology of the textbooks, the general performance of the classifiers with respect to the language level and the linguistic knowledge used for creating the model are studied. Based on the obtained results, the amount of extracted features plays an important role in training a classifier. Furthermore, the average best performance of the classifiers is extending the linguistic knowledge from syntactic patterns at proficiency level A (beginner) to all linguistic information at levels B (intermediate) and C (advanced).

keywords: machine learning, classification, feature, computational cognitive model, Persian learner

[1] Assistant Professor, Faculty of Linguistics, Institute for the Humanities and Cultural Studies, Tehran, Iran; M.Ghayoomi@ihcs.ac.ir

DOI: 10.22051/LGHOR.2021.32656.1354

Introduction

Language proficiency assessment is a fundamental step within the language learning process. This task should be done precisely to be used in fundamental decisions such as demanding for work or obtaining the study permission. To reach the goal, various exams are compiled to precisely evaluate the language skills, including listening, speaking, reading, and writing. To this end, Test Of English as a Foreign Language (TOEFL) and International English Language Testing System (IELTS) as the two widespread, major, distinguished tests are designed to evaluate the proficiency level of English. The Common European Framework of Reference for Languages (CEFR; Council of Europe 2001) is a framework of reference to provide a transparent, coherent, and comprehensive basis for the language teaching syllabuses and guidelines. This framework is used in Europe and also in other continents. The application of using this framework caused to know 3 groups of language proficiency levels, namely beginner (A), intermediate (B), and advanced (C). The levels are extended to six-point scale known as the basic user (A1 and A2), the dependent user (B1, B3), and the proficient user (C1, and C2). CEFR guideline has provided a complete description for each language level and defined the properties of each level.

Language learning procedure has also been studied from cognitive linguistics perspective (Belkhir, 2020). Matlin (2005, p. 2) defined "cognition" as a mental activity with various cognitive processes. This statement determines how wide the cognition is. One property of this field is to translate the CEFR language levels' properties into linguistic features. These features can be transformed in such a way to be used by a computer to simulate the language learning process and create a model. It should be noted that, as Macwhinney (2010) stated, in computational modeling, the whole process of language learning cannot be simulated but partly.

In this paper, we aim to determine the language proficiency level of a given written text automatically by using computational linguistics methods. This approach explores the general content properties of Persian learners' textbooks. The outcome of this research can be used to discover the latent teaching methodology in the Persian learners' textbooks and to check their major and minor focuses on a linguistic knowledge. This knowledge can be utilized to increase the content quality of the textbooks and to have a uniformed distribution of the linguistic knowledge for all levels. One additional application of creating such a model is predicting the language proficiency level of essays in the second language and discovering the linguistic knowledge that the learner is master at it.

The structure of the paper is as follows: in Section 2, we briefly overview the previous studies on developing methods to simulate the language learning process and to measure the language proficiency levels. The theoretical framework of second or foreign language learning is discussed in Section 3. Section 4 describes our proposed algorithm to determine the language proficiency level of a written text. In Section 5, the obtained results are reported and discussed. And finally, in Section 6, we conclude the paper.

Previous Studies

Two concerns are found in the literature, when one studies readability assessment. One is the application of studying the readability of a text for different topics, and the other one is the methodology how to assess the readability of a text.

There are a large number of researches that have focused on readability of a

Persian text from different perspective, for instance studying readability and linguistic properties of fictions (Khademizadeh & Vaezi, 2020; Vaezi et al., 2016), readability of school textbooks (Ghaderi Moghaddam & Sobhaninejad, 2016; Khodadady & Mehrazmay, 2017; Nazari et al., 2016; Shekari & Najareyan, 2012), readability of pharmaceutical brochures (Zarea Gavgani et al, 2018), readability of patients (Ahmadzadeh et al., 2014), readability of translation studies (Maftoon & Daghigh, 2001), readability of accounting standards (Sarvi et al., 2019), and readability of health information (Zeinali et al., 2019). In these studies, the readability property of various domains is studied. However, the research by Mohammadi and Khaste (2020) on Persian is the only research that used machinery methods to assess the readability of an open domain text collected through crowd-sourcing. In the data collection process, 400 people participated to collect 12780 texts. The data is labeled as easy, medium, and hard. In their proposed model, 5 classifiers were used to predict the difficulty level of the texts. To train the classifiers, vectors which contained statistical features, such as average sentence length, average word length, and linguistic patterns based on *n*-grams ($1 \leq n \leq 5$) extracted from word forms along with their part-of-speech tags, were used

Although there are studies such as Doró (2011) who used statistical analyses to predict educational success through language proficiency, ten Bosch et al. (2009) who proposed a computational model for human cognitive process to acquire a language, or Matusevych et al. (2013) who proposed a model to study the impact of construction priming and statistical distribution on learning a second language, the current research aims at predicting language proficiency level of a written Persian text automatically.

Luo et al. (2008) studied oral language proficiency through signal processing. The shadowing method was used such that the learners ought to repeat the utterance of the instructor immediately. Pronunciation of learners belonging to lower proficiency level had delays and errors and it could not match with the instructor's production.

de Wet et al. (2009) proposed a model to evaluate oral language proficiency and listening comprehension. To this end, a spoken dialogue system was developed. To evaluate the language proficiency, speech rate, pronunciation goodness as well as repeat accuracy were calculated. According to the results, speech rate provided a fair indication of oral proficiency.

Crossley et al. (2011) studied the application of lexical indices to predict the proficiency level of the language learners. In their study, a set of 1000 writing samples in second language was collected from 100 learners. The texts were categorized into three groups, namely beginner, intermediate, and advance. Wide range of lexical knowledge was extracted from the data. The extracted features were used to train a classification model to predict the language proficiency level. According to the results, 70% of the texts were correctly classified.

Pilán, Alfter et al. (2016) proposed a classification model and used a machine learning method to determine the language proficiency level of an essay written by a human learner and to enhance the model by utilizing reading passages in the language learning textbooks. Experimental results of the proposed model showed that incorporating features from the latter dataset boost the performance of the model significantly.

Pilán, Volodina et al. (2016) aimed at proposing a model to predict the language proficiency of Swedish. To this end, they used a classification model. One

main issue they studied was the data sparsity problem and using the data in second language that belonged to another domain to overcome this problem. In the model, they used statistical as well as linguistic features, including morphological, syntactical, and semantic features. These features were used to train the Support Vector Machine (SVM) classifier (Boser et al., 1992).

Yang et al. (2016) proposed a classification model to predict language level proficiency. They used linguistic cognitive properties to the model to improve the performance of their model.

Monaghan et al. (2017) explored the relations between word frequency, language exposure, and bilingualism in a computational model of reading.

Jung et al. (2019) used computational tools to predict the second language writing proficiency based on the learners' texts. In this study, it was found that linguistic features associated with text length and lexical complexity were the most important predictive elements of the writing quality.

McLean et al. (2020) used statistical methods to predict reading proficiency of the second language. In their study, the correlation between vocabulary knowledge and reading proficiency were calculated.

Theoretical Frameworks

There are three general language acquisition theories. One is the *nativist* approach that is rooted at the Chomskyan attitude towards language learning (Chomsky, 1965; 1968; 1980). Rationalists who follow this approach provide a formal representation of the language and they believe that language is instinct (Pinker, 1994) and there is an innate language learning capability in the brain known as universal grammar that is given to a language learner for free and tuning is required to fully acquire the language.

The other one is the *emergentist* explanation approach that is more functional and usage-based (MacWhinney, 1999). This approach rooted at this idea that the language structure emerges from the language use and it is not instinct (Evans, 2014). This attitude is nothing but patterns as a sequence of meaningful linguistic symbols that are induced by a child. Therefore, all complex cognitive activities as an abstract knowledge are acquired to construct one's language.

The third one is the *cognitive* approach within the constructivism theory that was proposed by two theorists Jean Peaget and Lev Vygotsky in 1930 and 1934, respectively (Sulistyowati, 2019). Constructivism is "an approach to learning that holds that people actively construct or make their own knowledge and that reality is determined by the experiences of the learner" (Elliott et al., 2000, p. 256). Constructivist item-based learning is proven to be children's early linguistic competence (MacWhinney, 2005; Tomasello, 2000); therefore, the language development process is from specific to general and it is organized around a concrete linguistic schema. For instance, to learn words, children experience various and rich contexts to construct the knowledge received from different contexts.

Undoubtedly each child acquires a language as a mother tongue language. But due to the living situation, (s)he might acquire additional language after the native language, called the second language. The second language can have a function in the place (s)he lives, but it should bear in mind that life situation may vary and there is a possibility that the person might learn a language for a very limited use in a classroom. This language is known as the foreign language. Theoretically, there is no difference between the second and foreign language

methods as Ellis (1997, p. 3) stated that the "second [language] is not intended to contrast with 'foreign' [language]. Whether you are learning a language naturally as a result of living in a country where it is spoken, or learning it in a classroom through instruction, it is customary to speak generically of 'second' language acquisition."

There are methods and approaches to teach a second or foreign language. Methods refer to prescriptions for a teacher and a learner on how to act during the teaching or learning process. Djigunović and Krajnović (2005) have briefly described the methods used for teaching or learning a second or foreign language. These methods that have been used for a long period of time include a) grammar translation, b) direct method, c) audio-lingual method, and d) the cognitive code learning.

Approaches are the theories on the nature of language and language learning. In addition to the language acquisition theories that were briefly described above, there are other approaches for teaching a second or foreign language, including Communicative Language Teaching (CLT), Task-based Language Learning (TLL), and Computer Assisted Language Learning (CALL).

In 1970s, CLT influenced language teaching (Brumfit & Johnson, 1979) such that language could be learnt through interaction. In this teaching approach, the mere attention to language structure was replaced by focus on the meaning and conveying information to one another (Widdowson, 1978).

In 1980s, CLT became mature and gradually the 'communicative' activity was replaced by a 'task'. Prabhu (1987) proposed a teaching approach called TLL. In TLL, learners are expected to do meaningful tasks using the target language, and much attention has been paid to meaning. Skehan (1998) and Robinson (2001) believed that, in TLL, learners make generalizations based on the forms in an attention-driven perspective. This approach fits with the constructivist item-based learning where the language develops from specific to general.

The first attempts to use CALL back to 1960s in the PLATO project. The development of this learning approach continued through progress in technology and the advent of microcomputers in the late 1970s (Marty, 1981). Levy and Stockwell (2006, pp. 249, 253) explored that "there is no one single TLL methodology" and CALL is both "interdisciplinary and multidisciplinary" in orientation. This idea persuaded Thomas and Reinders (2010, pp. 4-5) to find synergies and the common ground between TLL and CALL approaches. They introduced micro and macro processes in language learning. The micro-level strategy involves "practical teaching, task framework, design and evaluation decisions based on particular learning contexts" and the macro-level strategy involves "analysis and integrated task, syllabus and curriculum design". TLL contains both micro and macro processes, while CALL contains only the micro process. This attitude towards language learning brought technology into account and the term Technology-Enhanced Language Learning (TELL) was used in 1990s instead of CALL (Bush & Terry, 1997). Internet and Web-based applications introduced another term called E-learning, which is defined by UNESCO as a type of learning by means of Internet or multimedia. One advantage of E-learning is that the learner is self-dependent which activates the cognitive motivation.

One of the tasks for second or foreign language learning is learning the required vocabulary to achieve fluency in a language. There are two learning strategies, namely incidental (implicit) learning, and intentional (explicit) learning

(Postman & Keppel, 1969). In incidental learning, the vocabularies are learnt without placing the focus on a specific word to be learned (Paribakht & Wesche, 1999), while in intentional learning, the learners are aware in advance what is going to be learned. Uchihara et al. (2019) found that exposure to the target vocabulary in a second language, which was originally studied by Ebbinghaus (1964), for a certain number of times affects the likelihood of the vocabulary to be learned. This achievement means that repetition has an impact on learning. In repeated exposure method, words are learned in the diversity of contexts; as a result, the learners do not learn the word forms such as the verbs at once but generalized knowledge is obtained at a later learning stage (Tomasello, 1992). The repeated exposure method causes learners to find linguistic patterns through statistical learning and analogy to create a more abstract knowledge about the language (Tomasello, 2006). This idea explores the importance of statistical information in the cognitive approach of language learning. The statistical learning attitude towards language learning is totally ignored by Chomskyan linguistics because "any account which assigns a fundamental role to segmentation, categorization, analogy, and generalization" is rejected as "mistaken in principle" (Chomsky, 1975). However, there are a number of researches on the cognitive approach of the language learning that determine how a learner is sensitive to the statistical structure of their linguistic input (Aslin et al., 1998; Gomez & Gerken, 1999; Newport & Aslin, 2000; Saffran et al., 1996). These findings have made progress to propose computational models to simulate the language learning process. The computational models of cognitive process deepen understanding of how induction methods are used to learn a language. Pinker (1996, p. 13) explored that a mathematical learning model contains 4 parts: 1) properties of the language within the scope of the learner's acquisition capability; 2) instances that the computational model uses to learn; 3) the learning algorithm; 4) the criteria to evaluate the model and to make it possible to conclude how well the proposed algorithm works.

Within the cognitive item-based repeated exposure framework and the interdisciplinary property of CALL, in this paper, we propose a computational model to predict language proficiency level of Persian learners' textbooks by utilizing the linguistic and statistical properties of the language to build a statistical model and study the underlying methodology used for compiling the textbooks.

Proposed Algorithm

Our proposed algorithm to determine the language proficiency level of a text has 4 components, according to Pinker (1996): a) the data to build the model; b) the feature selection component to extract the required knowledge from the data; c) statistical classifiers as learners to use the features and to predict the language proficiency level of a text; and d) evaluation. The components are described in the following.

Data

The data used in our model is a collection of texts in textbooks for teaching Persian to non-Persian speakers. Various Persian learning textbooks at different levels, including the Basic (A), Intermediate (B), and Advanced (C) levels, are available that are listed below:

- **Source 1**: "Teaching the Persian Language" in basic, intermediate, and advanced levels by Samareh (1989; 2005a; 2005b; 2005c);

- **Source 2**: "The Persian Lesson for Foreign Persian Learners" in basic level by Poornamdariyan (1994);
- **Source 3**: "Series of Teaching the Persian Language" in basic, intermediate, and advanced levels by Zarghamiyan (1998; 2001a; 2001b);
- **Source 4**: "General Persian" in basic level by SaffarMoghaddam (2003);
- **Source 5**: "The Persian Language" in basic, intermediate, and advanced levels by Saffar Moghaddam (2008a; 2008b; 2008c; 2008d);
- **Source 6**: "Let's Learn Persian" in basic, intermediate, and advanced levels by Ghaffari et al (2004).

To collect this data as the corpus to be used in our model to extract the required statistical information, two people typed the print version of the selected passages in the books and provided the electronic format of the texts as a corpus. The selected passages were complete and coherent texts, mostly from the reading comprehension sections. The texts did not have a dialogue format, and the texts in the exercises were not used. Table 1 summarizes the general statistical information extracted from the target sources. It has to be emphasized that we assumed that the defined levels of the textbooks are correct and we utilized the levels as indicated by the labels to represent the proficiency levels.

Table 1
Detailed Statistical Information of Target Sources

Source	Level	Sentences	Tokens	Types	Lemmas
Source1	A	149	1657	656	591
	B	359	5310	2414	2106
	C	552	11190	4730	4260
Source2	A	272	2957	1585	1453
Source3	A	130	967	411	393
	B	373	3579	1714	1604
	C	776	10306	5000	4543
Source4	A	221	2828	1649	1580
Source5	A	88	2926	926	551
	B	135	2849	1653	1585
	C	209	4662	2431	2317
Source6	A	143	1269	699	672
	B	379	5470	3217	3049
	C	1633	28857	11698	10445

Comparing the extracted statistics of texts belonging to different levels in Table 2 reveals that, as the level goes higher, the number of sentences, word forms, and lemmas to compile the texts increase. Although the number of texts for different levels is almost balanced, Level C contains a large number of sentences, word forms, and lemmas.

Table 2
Summary of Statistical Information of Target Sources

Level	Texts	Sentence	Tokens	Types	Lemmas
A	88	1004	10604	1968	1645
B	84	1247	17206	3499	2863
C	87	3171	55004	8767	7112

Feature Selection

Basic Information. The collected raw corpus has to be normalized and tokenized to acquire reliable results. Then, the data has to be analyzed linguistically. The linguistic information that is added to the data varies from phonological, to morphological, syntactical, and semantic information.

In the normalization process, character codes for the Arabic letters Yeh "ي" and Kāf "ك" were replaced with the equivalent Persian letters as "ی" and "ک". In the tokenization process, space was added between the letters, punctuation marks, and numbers as a word boundary to recognize each token (word). These two tasks were done automatically. Moreover, the extra white space was replaced by a pseudo-space to resolve the multi-unit token problem (Ghayoomi, 2018). This task was done semi-automatically.

Five linguistic analyses were run automatically on the data. The first one was assigning phonological patterns to each word form. To this end, we used the productive lexicon developed by Eslami et al. (2004). This word list contains lemmas and their phonological patterns. In our research, we drove the phonological pattern for each word form from the lemmas, and then assigned the pattern to the target words.

The two next analyses were lemmatization and Part-Of-Speech (POS) tagging. To these ends, we used the Marmot tool (Müller et al., 2013) for POS tagging and the Lemming tool (Müller et al., 2015) for lemmatization. These two toolkits were adapted for Persian by Ghayoomi (2019a) who compared the performance of different tools to lemmatize and POS tag Persian words. Based on the reported experimental results, the toolkits Marmot and Lemming outperformed the other toolkits. These results persuaded us to use these tools to annotate our corpus. To train the tools with the Persian data, we used the Bijankhan Corpus (Bijankhan, 2004) that has already been POS tagged semi-automatically and Ghayoomi (2019b) lemmatized the corpus semi-automatically. The tag set that was used in the Bijankhan Corpus is fine-grained and it contains 586 POS tags. Ghayoomi (2012) standardized the POS tags according to the Multi-Text East standard. In this standard, the length of each tag became fixed with respect to the main category of the tag; furthermore, specific information was defined in certain positions. The standardized data was used for training the tools.

The annotated data after lemmatization and POS tagging was the input to two other tools to do the syntactic analyses and to provide the parse tree of sentences. To this end, constituent parsing and dependency parsing of the data were performed. The toolkit used for constituency parsing was the statistical Stanford Parser (Klein & Manning, 2003) adapted to Persian by Ghayoomi (2013). To train this toolkit, the Persian constituency treebank developed by Ghayoomi (2012) was used. This treebank was developed within the linguistic framework of Head-driven Phrase Structure Grammar (Pollard & Sag, 1994). One property of this treebank is

that four types of head- dependent relations are defined in the tree analysis of each sentence and the trees are decorated with this additional information. The treebank contains 1024 sentences from Bijankhan Corpus. The dependency parser used in our research was the Mate Parser (Bohnet, 2009). The dependency parser was trained with the dependency treebank developed through conversion from the constituency treebank by Ghayoomi and Kuhn (2014). The dependency treebank contained 49 unique dependency relations. In the next step, the data standardization was required. When all linguistic analyses were done, all information was collected and standardized according to the CoNLL data format (Ghayoomi, 2020).

The CoNLL data format was proposed by the Conference on Natural Language Learning (CoNLL) in 2006. In this data format, each word appears in one row and the related information is defined in columns separated by a tab delimiter. The sentence border is determined by an empty line. Table 3 shows an example how the data from various sources is structured in 9 columns according to the CoNLL format.

Table 3

Sample of the Organized Data based on the CoNLL Format

Sent. ID	Word ID	Word Form	Phonological Pattern	Lemma	POS tag	Dep. Relation	Dep. Type	Constituency Tree
1	1	هادی	CVCV	هادی	Nasp---	5	NSUBJ	(ROOT (S (VPS (Nasp--- *)
1	2	کارمند	CVCCVCCV	کارمند	Ncsp--z	1	NN	(VPC (NPC (Ncsp----z- *)
1	3	اداره	CVCVCVCV	اداره	Ncspk-z	5	COPCOMP	(NPC (Ncspk---z- *)
1	4	پست	CVCC	پست	Ncsp---	3	NN	(Ncsp------ *)))
1	5	است	CVCC	بودن	Vpykshs----	0	ROOT	(Vpyk-shs------ *))
1	6	.	.	.	Oe	5	PUNC	(Oe *)))

In this table, the following information is available:

- the sentence ID in a text;
- the word ID in a sentence;
- the word form;
- the phonological pattern of the word form;
- the lemma of the word form;
- the POS tag of the word form in the sentence;
- the ID of the head word on which the dependency tree analysis depends;
- the type of the dependency relation between the head and the dependent words;
- the constituency tree analysis in the CoNLL 2011 format (Ghayoomi, 2020) in such a way that the nodes in the tree are drawn till the target word form is met.

According to the available information described above, the basic statistical information was extracted from the annotated data related to the words and the whole text. This information for each word contains:

- the number of characters in each word;
- the number of syllables in each word;
- the number of the CVCC phonological pattern in each word;
- the number of the CVC phonological pattern in each word;

- the number of the CV phonological pattern in each word;
- the number of constituency nodes in the tree analysis related to each word;
- the number of named entities.

Training Features. Based on the basic statistical information mentioned in the previous section and the annotated data, 37 features were extracted from the data and represented as a 37- dimension vector to train the supervised machine learning models. We ought to bear in mind that the labels assigned to the vectors were the original textbooks' proficiency levels.

The features were categorized into 5 groups. The statistical information in the vectors was extracted from both the raw corpus and the annotated corpus that contained phonological, morphological, syntactical, and semantic labels. In the vector representation, all features were normalized according to the word, sentence, or the text. Therefore, the relative frequency rather than the absolute frequency was used to construct the vector in order to reduce the negative impact of the text length.

Before describing the features, it is necessary to compare the data we utilized in our research with the data that Mohammadi and Khasteh (2020) used. In our study, we used various linguistic information in addition to statistical information extracted from the corpus. This property made the features much richer than the features used by Mohammadi and Khasteh (2020) who used merely statistical information of word forms and POS tags. Moreover, we used the content of textbooks that are used for teaching Persian to non-native speakers. This property of the corpus controlled the data accurately in terms of lexicon, syntactic and semantic complexities in the texts; while the dataset developed by Mohammadi and Khasteh (2020) through crowd-sourcing was labeled by using people's intuition without making any scientific justification to assign a difficulty label to a text. In addition, if their labeled data is accepted, no balance was found in their developed data for different levels (54% as an easy text, 32% as a medium text, and 14% as a hard text). The weak point of their data is the imbalanced data problem, because dominance on a specific level misleads classifiers. However, we used a sort of balanced data in terms of the number of documents to train the classifiers as reported in Table 2 (34% for the level A, 32% for the level B, and 34% for the level C).

Statistical Information from Raw Corpus. The statistical feature set extracted from the raw corpus contained the following information:

(a) the average word length;
(b) the average sentence length;
(c) lexical diversity which is the ratio of word types over word tokens;
(d) the ratio of word types from the first 100 word tokens of a text over word tokens in a text;
(e) the ratio of words (unigrams) with frequency 1 over word tokens in a text;
(f) the ratio of words with frequency 1 over word tokens with frequency above 1;
(g) the ratio of word bigrams (sequence of two words) with frequency 1 over the total number of word bigrams in a text;
(h) the ratio of word bigrams with frequency 1 over word bigrams with frequency above 1;
(i) the ratio of word trigrams (sequence of three words) with frequency 1 over the total number of word trigrams in a text;
(j) the ratio of word trigrams with frequency 1 over word trigrams with

frequency above 1.

Statistical Information from Phonological Annotation. The extracted statistical feature set based on phonological annotation of the corpus contained the following information:

(a) the average syllable of the words in a text;
(b) the ratio of one-syllable words over word tokens;
(c) the ratio of two-syllable words over word tokens;
(d) the ratio of three-syllable words and above over word tokens;
(e) the ratio of the CVCC syllable pattern over the total number of syllables;
(f) the ratio of the CVC syllable pattern over the total number of syllables;
(g) the ratio of the CV syllable pattern over the total number of syllables;
(h) the ratio of one syllable words in the first 150 word tokens of a text over the total number of syllables;
(i) the ratio of three-syllable words and above in the first 100 word tokens of a text over the total number of syllables.

Statistical Information from Morphological Annotation. The extracted statistical feature set based on morphological annotation of the corpus contained the following information:

(a) the ratio of word lemmas over word tokens;
(b) the ratio of word lemmas over word types;
(c) the ratio of word lemmas with frequency 1 over word tokens;
(d) the ratio of word lemmas with frequency 1 over word tokens with frequency above 1.

Statistical Information from Syntactic Annotation. The extracted statistical feature set based on syntactic annotation of the corpus contained the following information:

(a) the ratio of functional words over word tokens;
(b) the ratio of functional words with frequency 1 over word tokens;
(c) the ratio of functional words over content words;
(d) the ratio of content words over word tokens;
(e) the ratio of content words with frequency 1 over word tokens;
(f) the ratio of content words over functional words;
(g) the ratio of POS bigrams with frequency 1 over the total number of POS bigrams in a text;
(h) the ratio of POS bigrams with frequency above 1 over the total number of POS bigrams in a text;
(i) the ratio of POS trigrams with frequency 1 over the total number of POS trigrams in a text;
(j) the ratio of POS trigrams with frequency above 1 over the total number of POS trigrams in a text;
(k) the ratio of dependency relation types over the total number of dependency relations in a text;
(l) the ratio of clause dependency relations over the total number of dependency relations in a text;
(m) the ratio of the number of nodes in constituency constructions over word tokens in a text;

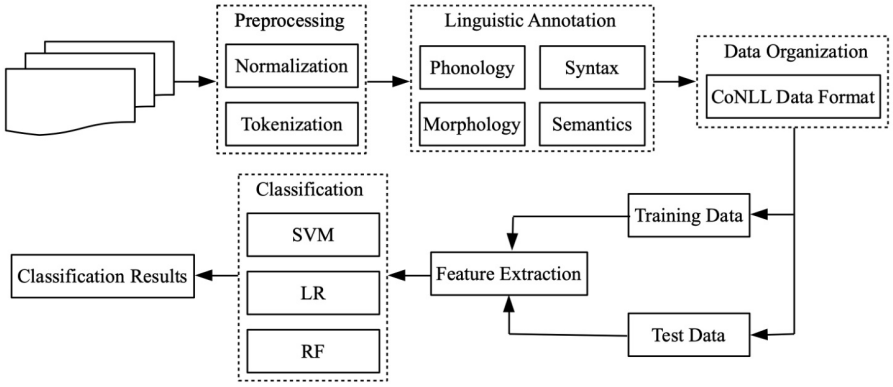
Statistical Information from Semantic Annotation. The extracted statistical feature set based on semantic annotation of the corpus contained the ratio of the named entities to word tokens in a text.

Our Proposed Algorithm

Figure (1) shows the architecture of our model. As can be seen in the figure, after data collection from Persian learners' textbooks, the corpus is pre-processed such that the texts are normalized to uniform the codes, and then the data is tokenized to identify each word. The cleaned data is linguistically annotated at the phonological, morphological, syntactical, and semantic levels. The annotated data from various levels should have a flat representation. To this end, we used the CoNLL data. A sample of this data structure is already shown in Table 3. We used Python programming language to develop our model.

Figure 1

Architecture of Our Proposed Model



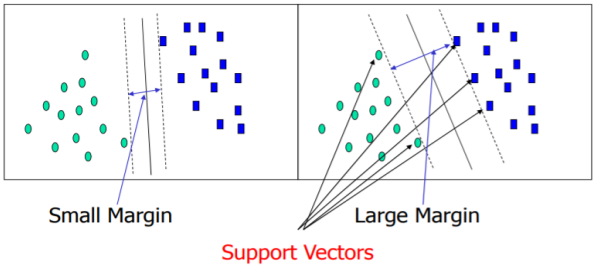
The prepared data that contains the labels of the proficiency levels is divided into two sets. The first division contains 90% of the data and it is used as training data to create the statistical model by a classifier. To this end, features are extracted from this data and they are represented as vectors to be used by the classifier. The second division of the data that contains 10% of the total data is used as test data to evaluate the performance of the classifier. For evaluation, we assume that the test data has no label and this data is given to the classifier to label. The output of the model on the test data and the original test data are compared to calculate the performance of the classifier.

Considering the nature of supervised learning techniques, the main two steps that have to be taken into the consideration are the feature engineering and the learning method. For the learning method, we utilize three algorithms that benefit from discriminative models, namely SVM (Boser et al., 1992), Logistic Regression (LR; Cramer, 2002), and Random Forest (RF; Breiman, 2001). One main property of discriminative models is that they use the inferred knowledge from a set of observed data. Although it seems that the three algorithms belong to a family, SVM works based on one single best margin with minimum risk of error, LR uses different weighted boundaries to make a near optimum decision, and RF uses stochastic discrimination over decision trees. Although deep neural networks have achieved state-of-the-art results in supervised learning, we cannot benefit from the deep neural approaches due to the small amount of the available data in our task. Therefore, three classifiers, namely SVM, LR, and RF, are used in our model to compare different sets of features and to study their impact on predicting language proficiency levels. Due to the shortage of amount data, no tuning for parameter

optimization is done and the default parameters in the classifiers are used. We use the scikit-learn[1] library in Python to call the classifiers into our code. The classifiers are briefly described in the followings.

SVM is a supervised machine learning method that uses the training data to build a model to assign a new instance to one or the other categories. To build the model, the data is represented in a vector space model. To make the distinctions, a hyperplane with maximum margin is used to create two subspaces. Figure (2) represents how the model decides between different possible hyperplanes based on their margin. The best hyperplane is the one with a large margin (James et al., 2013, pp. 337-342).

Figure 2
Support Vector Machine



LR, as another supervised machine leaning method utilized for classification, uses conditional probability assumptions that rely on the underlying data distribution represented as vectors. The probability score is a number between 0 to 1 (Indurkhya & Damerau, 2010, pp. 194-196).

RF (Breiman, 2001) consists of multiple random decision trees. While making the decision at each node tree, features are randomly selected to generate the best data split. In this classification model, feature selection has the most important contribution on the class probability. This means informative features confuse the model and they should be removed before fitting a classifier. However, redundant features reduce the importance of the features.

Evaluating the Algorithm

To evaluate the performance of our proposed algorithm, we use two evaluation metrics. One is calculating the accuracy of the model in Equation (1) for the total performance of the algorithm:

(1)

$$Accuracy = \frac{number\ of\ correctly\ predicted\ instances}{total\ \ number\ of\ instances}$$

The other evaluation metric is calculating F-measure in Equation (2) that is a harmonic mean of precision and recall proposed by van Rijsbergen (1979):

(2)

$$F - measur = \frac{(1+\beta) \times P \times R}{P + R}$$

[1] https://scikit-learn.org/stable/

where P is precision, R is recall, and β is a weighting parameter. If $\beta > 1$, more weight is assigned to recall, and in case $\beta < 1$, more weight is assigned to precision. If $\beta = 1$, precision and recall are considered equally. Equations (3) and (4) compute precision and recall, respectively:

(3)

$$P = \frac{number\ of\ correctly\ predicted\ instances\ for\ level\ X}{number\ of\ predicted\ instances\ for\ level\ X}$$

(4)

$$r = \frac{number\ of\ correctly\ predicted\ instances\ for\ level\ X}{number\ of\ instances\ for\ level\ X\ of\ gold\ data}$$

where X is one of the proficiency levels A, B, or C. After calculating precision and recall for each language level, we calculate the average precision and recall for all of the levels.

Since the total amount of data that we use in our research is not much, we use 10-fold cross validation method to evaluate our model; that is, the total amount of data is divided into 10 folds and in each round of experiments, one fold (10% of the data) is considered as the test data, and the rest (90% of the data) as the training data. The average of the obtained results, for both accuracy and F-measure, can be considered as the performance of the model.

Results and Discussions

This research aims at proposing a computational linguistic model to predict language proficiency level and to explore the general properties of the levels. To reach the goal, we configured different feature sets defined in Section 4.2.2 to train the classifiers for different levels. To make the comparison possible and to represent the superiority of the rich features to train a model, we require a baseline. To this end, we use Term Frequency-Inverse Document Frequency (TF-IDF) which represents how relevant a word to a document is in a collection of documents (Salton et al, 1975).

We proposed learning scenarios with different feature sets. The first learning scenario contained all available features that are the mixture of both statistical and linguistic features. The second and third learning scenarios contained statistical information and the combination of all linguistic information respectively. The fourth to seventh learning scenarios contained individual linguistic information to build separate models, including phonetics, morphology, syntax and semantics. After training the classifiers, the test data was given to the classifiers.

The confusion matrix of the labeled data for different classifiers belonging to different levels is reported in Table 4. In this table, the predicted labels of each classifier trained with different feature sets are compared to the gold data of a specific language level. As an example, the first row of Table 4 should be read in such a way that the gold data has 88 texts which belong to the proficiency level A. If the TF-IDF feature set is used, the SMV classifier assigns the predicted label C to all data. If the Statistical and Linguistic Features are used, the SVM classifier predicts the level A for 20 texts, the level B for 42 texts, and the level C for 26 texts out of 88 texts. As a result, 20 texts are labeled correctly, and the rest of 68 texts are labeled incorrectly. Using different feature sets causes the classifiers to behave differently.

Table 4
Confusion Matrix for Labeling of Classifiers for All Levels

Classifier	Level	Gold	TF-IDF			Statistics, Linguistics			Statistics			Linguistics			Phonetics			Morphology			Syntax			Semantics		
			A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C
SVM	A	88	0	0	88	20	42	26	28	31	29	31	48	9	16	65	7	24	31	33	40	32	16	36	22	30
	B	84	0	0	84	11	14	59	7	13	64	9	35	40	11	34	39	22	25	37	16	19	49	27	14	43
	C	87	0	0	87	14	5	68	12	10	65	1	6	80	3	2	82	9	10	68	2	2	83	16	9	62
LR	A	88	46	0	42	65	20	3	66	17	5	62	22	4	71	14	3	61	2	25	59	23	6	55	0	33
	B	84	43	0	41	21	44	19	22	43	19	21	50	13	35	20	29	60	1	23	23	47	14	33	1	50
	C	87	67	0	20	2	14	71	1	13	73	5	11	71	6	5	76	24	3	60	3	16	68	22	6	59
RF	A	88	37	0	51	66	19	3	66	17	5	69	16	3	68	17	3	37	27	24	67	16	5	53	9	26
	B	84	35	0	49	29	41	14	21	44	19	30	40	14	34	35	15	22	43	19	31	38	15	36	5	43
	C	87	58	0	29	1	9	77	1	11	75	3	9	75	3	8	76	9	17	61	0	18	69	21	12	54

Based on the number of labeled data in Table 4 for different classifiers and Equations (1) to (4), we evaluated the performance of the classifiers, using accuracy and F1 metrics. Table 5 reports the average performance of the models using 10-fold cross-validation method. In our experiments, we used TF-IDF as the basic feature set to compare the performance of the models that used different feature sets. As it is obvious from the results, the malperformance of the classifiers was obtained when trained with TF-IDF feature set. In general, the SVM classifier created the worst model, and the RF classifier outperformed the other models using features other than TF-IDF. This achievement determines that previous words (the history) play a very important role in the learning process. However, the SVM classifier in comparison to the LR and RF classifiers performed the best when using TF-IDF features set. The result shows that, for this classifier, history has a negative impact on predicting labels.

Comparing each classifier based on the feature sets that were used for training, the model created by SVM and LR classifiers performed the highest when a combination of linguistic information was used. Additionally, RF performed the best using statistical information. However, RF performed slightly worse than the model using either combination of linguistics information or combination of statistical and linguistic. This result determines that additional information could not improve the classifier's performance.

We further evaluated the performance of the classifiers using different features for each proficiency level. The results are reported in Table 6. While the SVM model performed the best when linguistic information was used in Table 5, the level C obtained the highest results among the three levels in Table 6. We further observed that using syntactic information for the level A in the SVM model outperformed using other feature sets at this level. The SVM model that used all linguistic information performed well for the level B; and using the phonetic information to build the SVM model caused to perform the best for the level C.

Comparing the performance of the LR models, the model utilized the linguistic information to build the model obtained the highest results in Table 5, and

the level C for this feature set obtained the highest result among the three levels in Table 6. We further observed that using statistical information for the level A in the LR model outperformed using other feature sets at this level. Additionally, the LR model that utilized all linguistic information performed the best for the levels B and C.

Comparing the performance of the RF models for different levels, the utilized statistical information in Table 5 that obtained the highest results performed the highest for the level C in Table 6. We further observed in Table 6 that using statistical information for the levels A and B in the RF model outperformed the models that utilized other features. Moreover, the RF model that used both statistical and linguistic information performed the best for the level C.

Table 5
Performance of the Classifiers Using Different Feature

Classifier	Evaluation Metric	TF-IDF	Statistics Linguistics	Statistics	Linguistics	Phonetic	Morphology	Syntax	Semantic
SVM	Accuracy	33.54	39.35	40.85	56.32	50.94	45.12	54.82	43.20
	F1	16.31	37.57	40.46	57.95	51.33	44.91	54.99	41.94
	Recall	33.33	39.60	41.24	56.69	51.32	46.00	55.37	42.16
	Precision	11.18	35.75	39.70	59.26	51.34	43.86	54.61	41.73
LR	Accuracy	25.46	69.51	70.28	70.62	64.48	47.09	67.17	44.42
	F1	13.41	69.56	70.45	70.80	64.31	40.01	68.06	37.45
	Recall	33.33	69.58	70.63	70.92	64.53	47.60	68.26	44.92
	Precision	8.49	69.55	70.27	70.68	64.09	34.50	67.88	32.12
RF	Accuracy	25.46	71.05	71.42	71.05	69.08	54.37	67.20	43.28
	F1	13.41	70.57	71.96	71.14	68.44	55.29	66.50	38.33
	Recall	33.33	71.06	72.23	71.62	68.95	55.91	66.91	42.48
	Precision	8.49	70.07	71.68	70.66	67.94	54.68	66.09	34.92

Table 6
F-measure of the Classifiers Using Different Features for each Individual Label

Classifier	TF-IDF			Statistics, Linguistics			Statistics			Linguistics			Phonetic			Morphology			Syntax			Semantic		
	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C
SVM	0	0	50.29	30.08	19.31	56.67	41.48	18.84	53.06	48.06	40.46	74.07	27.12	36.76	76.28	33.57	33.33	60.44	54.79	27.74	70.64	43.11	21.71	55.86
LR	37.70	0	21.05	73.86	54.32	78.89	74.58	54.78	79.35	70.45	59.88	81.14	71.00	32.52	77.95	52.36	2.22	61.54	68.21	55.29	77.71	55.56	2.20	51.53
RF	33.94	0	26.85	71.74	53.59	85.08	75.00	56.41	80.65	72.63	53.69	83.80	70.47	48.61	83.98	47.44	50.29	63.87	72.04	48.72	78.41	53.54	9.09	51.43
Average	23.88	0	32.73	58.56	42.41	73.55	63.69	43.34	71.02	63.71	51.34	79.67	56.2	39.3	79.4	44.46	28.61	61.95	65.01	43.92	75.59	50.74	11	52.94

Additionally, we investigated which feature set has impact on the performance of a classifier with respect to the language proficiency level. In this analysis, we calculated the average performance of the classifiers to focus on the latent teaching methodology used for compiling the textbooks. The results are reported in the last row of Table 6. The average best performance of the learners for the level A is achieved by using syntactic information; i.e. the syntactic information is the main focuses of the textbooks at this level to make the learners aware of the basic constructions. The average best performance of the learners for the levels B and C is achieved by using linguistic information. This result determines that a great amount of attention is given to linguistic information at levels B and C. Among the linguistic information, syntactic and phonetic information play the most important role for levels B and C, respectively. One reason is extracting various features for these two linguistic components from the annotated data; i.e., the higher the number of features to be extracted, the better a classifier learns about the properties of the language.

Concluding Remarks

In this paper, we proposed a computational linguistic model to detect the language proficiency level of a given text and to label it automatically by using machine learning methods. To reach the goal, we defined sets of features including statistical and linguistics. The linguistic features contained phonological, morphological, syntactic, and semantic features. Additionally, we used TF-IDF feature set as the baseline to compare the performance of the models. Our proposed models that used statistical and/or linguistic features outperformed the baseline. The features were extracted from a corpus developed from 6 Persian learners' textbooks that belonged to the beginner, intermediate, and advanced levels. The collected data was divided into subsets to train and test the classifiers. Based on the results, the

model created by the RF classifier performed the best using statistical features. This determined that enriching the model and providing more information does not guarantee to achieve the best performance. But this was not a global finding because it totally depends on the learning algorithm of the classifier, because the linguistic information caused the SVM classifier to perform the best.

We studied the performance of the classifiers with respect to the language proficiency level and the linguistic knowledge used to create the model. One property of the texts at the level A was paying much attention to the syntactic constructions. The general property of the texts at the levels B and C was using all linguistic information to compile the textbooks.

The outcome of this research can be used to check major and minor focus of the Persian learners' textbooks on linguistic knowledge and to increase the quality of the textbooks by utilizing uniformed distribution of linguistic knowledge for all levels. A sample for minor focus of the textbooks is the low performance of the classifiers for the level B when the semantic feature is utilized and the classifiers of this level had a malperformance compared to the levels A and C.

Acknowledgement

This research is funded by Iran National Science Foundation (INSF) for the research proposal number 97000696.

References

Ahmadzadeh, K., Khosravi, A., Arastoopoor, S., & Tahmasebi, R. (2014). Assessing the readability of patient education materials about diabetes available in Shiraz Health Centers. *Iranian Journal of Medical Education, 14*(8), 661-667. http://ijme.mui.ac.ir/article-1-3157-en.pdf

Aslin, R., Saffran, J., & Newport, E. (1998). Computation of conditional probability statistics by 8-month old infants. *Psychological Science*, *9*, 321-324. https://doi.org/10.1111/1467-9280.00063

Belkhir, S. (2020). Cognition and language learning: An introduction. In S. Belkhir (Ed.), *Cognition and language learning* (pp. 1-12). Cambridge Scholars Publishing.

Bijankhan, M. (2004). The role of corpora in writing a grammar: Introducing a software. *Journal of Linguistics*, *19*(2), 48-67.

Bohnet, (2009). Efficient parsing of syntactic and semantic dependency structures. In *Proceedings of the 13th conference on computational natural language learning: Shared task* (pp. 67-72). Association for Computational Linguistics. https://www.aclweb.org/anthology/W09-1210.pdf

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the 5th annual workshop on computational learning theory* (pp. 144-152). https://doi.org/10.1145/130385.130401

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Brumfit, C., & Johnson, K. (1979). *The communicative approach to language teaching*, Oxford University Press.

Bush, M., & Terry, R. (1997). *Technology-enhanced language learning*, National Textbook Company.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. The MIT Press.

Chomsky, N. (1968). *Language and Mind*. Harcourt Brace Jovanovich.

Chomsky, N. (1975). *Reflections on language*. Pantheon Books.

Chomsky, N. (1980). Rules and representations. *Behavioral and Brain Sciences*, *3*, 1-61. https://doi.org/10.1017/S0140525X00001515

Cramer, J. S. (2002). *The origins of logistic regression*. Technical Report (pp. 167-178). Tinbergen Institute. https://papers.tinbergen.nl/02119.pdf

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2011). Predicting the proficiency level of language learners using lexical indices. *Language Testing*, *29*(2), 240-260. https://doi.org/10.1177/0265532211419331

de Wet, F., Van Der Walt, C., & Niesler, T. R. (2009). Automatic assessment of oral language proficiency and listening comprehension. *Speech Communication*, *52*, 864-874. https://doi.org/10.1016/j.specom.2009.03.002

Djigunović, J. M., & Krajnović, M. M. (2005). Language teaching methodology and second language acquisition. In V. Muhvic-Dimanovski & L. Socanac (Eds.), *Encyclopedia of life support systems*, (pp. 394-417). Eolss Publishers Co. Ltd.

Doró, K. (2011). English language proficiency and the prediction of academic success of first-year students of English. *UPRT 2010: Empirical studies in English applied linguistics* (pp. 173-186). Lingua Franca Csoport. http://publicatio.bibl.uszeged.hu/11049/1/Doro%202011%20Language%20proficiency%20and%20academic%20success.pdf

Ebbinghaus, H. (1964). *Memory: A contribution to experimental psychology*. (H. A. Ruger & C. E. Bussenius, Trans.). Dover Publications. (Original work published 1885). https://doi.org/10.5214/ans.0972.7531.200408

Elliott, S. N., Kratochwill, T. R., Littlefield, C., J., & Travers, J. (2000). *Educational psychology: Effective teaching, effective learning (3rd Ed.)*. McGraw-Hill College.

Ellis, R. (1997). *Second language acquisition*. Oxford University Press.

Eslami, M., Mosavi Atashgah, M., Alizadeh Lamjiri, S., & Zandi, T. (2004). Persian productive lexicon. In *Proceedings of the 1st workshop on the Persian language and computer*, University of Tehran.

Evans, V. (2014). *The language myth: Why language is not an instinct*. Cambridge University Press.

Ghaderi Moghaddam, M. E., & Sobhaninejad, M. (2016). Validation methods to measure textbooks readability. *Research in Curriculum Planning*, *13*(21), 44-55.

Ghaffari, M., Mahmoodi Bakhtiyari, B., & Zolfaghari, H. (2004). *Let's learn Persian* (Volumes 1-3). Madreseh Publication.
https://jsr-e.khuisf.ac.ir/article_534415_65a3945c9994bc90c81c23ab0eacfaf7.pdf?lang=en

Ghayoomi, M. (2012). Bootstrapping the development of an HPSG-based treebank for Persian. *Linguistic Issues in Language Technology*, *7*(1).

Ghayoomi, M. (2013). Word clustering for Persian statistical parsing. In H. Isahara, & K. Kanzaki, (Eds.), *Advances in natural language processing*, (pp. 126-137). Springer.
https://doi.org/10.1007/978-3-642-33983-7_13

Ghayoomi, M. (2018). The problem of multi-words in syntactic processing of Persian. In *Proceedings of the fourth nation conference on computational linguistics* (pp. 11-40). Neviseh Parsi Publications.

Ghayoomi, M. (2019a). *Studying issues for automatic processing of the Persian language on lemmatization, part-of-speech tagging, and parsing: Developing a software using machine learning methods*. Technical Report. Tehran, Iran.

Ghayoomi, M. (2019b). Transition from rule-based to statistical lemmatization in Persian. In *Proceedings of the 5th national conference on computational linguistics* (pp. 57-86). Neveeseh Parsi Publications.

Ghayoomi, M. (2020). Structuring multilayer linguistic analyses in linguistic corpora. In F. Ghatreh & Sh. Modarres Khiabani, (Eds.), *Word by word of life: Festschrift for professor Vida Shaghaghi* (pp. 287-312). Neveeseh Parsi Publications.

Ghayoomi, M., & Kuhn, J. (2014). Converting an HPSG-based treebank into its parallel dependency-based Treebank. In *Proceedings of the 9th international conference on language resources and evaluation* (pp. 802-809). Reykjavik, Iceland.
http://www.lrec-conf.org/proceedings/lrec2014/pdf/441_Paper.pdf

Gomez, R., & Gerken, L. (1999). Artificial grammar learning by one-year-olds leads to specific and abstract knowledge. *Cognition*, *70*, 109-135.
https://doi.org/10.1016/S0010-0277(99)00003-7

Indurkhya, N., & Damerau, F. J. (2010). *Handbook of natural language processing*. Chapman & Hall.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in R*. Springer.

Jung, Y. J., Crossley, S., & McNamara, D. (2019). Predicting second language writing proficiency in learner texts using computational tools. *Journal of Asia TEFL*, *16*(1), 37-52.
https://doi.org/10.18823/asiatefl.2019.16.1.3.37

Khademizadeh, S., & Vaezi, M. R. (2020). Evaluating readability of Persian fictions selected by flying Turtle the Iranian award. *Publishing Research Quarterly*, *36***,** 116-128.
https://doi.org/10.1007/s12109-019-09705-0

Khodadady, E., & Mehrazmay, R. (2017). Evaluating two high intermediate EFL and ESL textbooks: A comparative study based on readability indices. *Sociology International Journal*, *1*(3), 93-102.
https://doi.org/10.15406/SIJ.2017.01.00016

Klein, D., & Manning, C. D. (2003). Accurate unlexicalized parsing. In *Proceedings of the 41st meeting of the association for computational linguistics* (pp. 423-430).
https://doi.org/10.3115/1075096.1075150

Levy, M., & Stockwell, G. (2006). *CALL dimensions: Options and issues in computer*

assisted language learning. Lawrence Erlbaum Associates.

Luo, D., Minematsu, N., Yamauchi, Y., & Hirose, K. (2008). Automatic assessment of language proficiency through shadowing. In *Proceedings of 6th international symposium on Chinese spoken language processing* (pp. 41-44). https://doi.org/10.1109/CHINSL.2008.ECP.22

MacWhinney, B. (1999). *The emergence of language*. Lawrence Erlbaum Associates.

MacWhinney, B. (2005). Item-based constructions and the logical problem. In *Proceedings of the workshop on psychocomputational models of human language acquisition* (pp. 53-68). Ann Arbor, Michigan. https://doi.org/10.3115/1654524.1654531

MacWhinney, B. (2010). Computational models of child language learning: An introduction. *Journal of Child Language*, *37*(3), 477-485. https://doi.org/10.1017/S0305000910000139

Maftoon, P., & Daghigh, M. (2001). Metric of determining readability of translated texts from English into Persian. *Humanities Bulletin*, *29*, 61-80. https://www.sid.ir/fa/journal/ViewPaper.aspx?id=27487

Marty, F. (1981). Reflections on the use of computers in second language acquisition. *System, 9*(2), 85-98. https://eric.ed.gov/?id=ED218932

Matlin, M. W. (2005). *Cognition*. John Wiley and Sons.

Matusevych, Y., Alishahi, A., & Backus, A. (2013). Computational simulations of second language construction learning. In *Proceedings of the workshop on cognitive modeling and computational linguistics* (pp. 47-56). Sofia, Bulgaria. Association for Computational Linguistics. https://www.aclweb.org/anthology/W13-2606.pdf

McLean, S., Stewart, J., & Batty, A. O. (2020). Predicting L2 reading proficiency with modalities of vocabulary knowledge: A bootstrapping approach. *Language Testing*, *37*(3), 389-411. https://doi.org/10.1177/0265532219898380

Mohammadi, H., & Khasteh, S. H. (2020). A machine learning approach to Persian text readability assessment using a crowd-sourced dataset. In *Proceedings of the 28th Iranian conference on electrical engineering*, University of Tabriz. https://doi.org/10.1109/ICEE50131.2020.9260933

Monaghan, P., Chang, Y. N., Welbourne, S., & Brysbaert, M. (2017). Exploring the relations between word frequency, language exposure, and bilingualism in a computational model of reading. *Journal of Memory and Language*, *93*, 1-21. https://doi.org/10.1016/j.jml.2016.08.003

Müller, T., Cotterell, R., Fraser, A., & Schütze, H. (2015). Joint lemmatization and morphological tagging with lemming. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 2268-2274). Lisbon, Portugal. Association for Computational Linguistics. https://www.aclweb.org/anthology/D15-1272.pdf

Müller, T., Schmid, H., & Schütze, H. (2013). Efficient higher-order CRFs for morphological tagging. In *Proceedings of the 2013 conference on empirical methods in natural language processing* (pp. 322-332). Seattle, Washington, USA. Association for Computational Linguistics. https://www.aclweb.org/anthology/D13-1032.pdf

Nazari, F., Farhadpour, M. R., & Soleymani, E. (2016). Measure the readability of the Persian text of the 'Lets know More' section of the Quran book for the grades two, three, and four of elementary school based on the Flash-Diani and Galing-Diani formulas. *Quarterly Journal of Knowledge and Information Management*, *3*(3), 85-92. http://lib.journals.pnu.ac.ir/article_4415_f2b05f84f03592edc72327a8a72ec55b.pdf?lang=en

Newport, E., & Aslin, R. (2000). Innately constrained learning: Blending old and new

approaches to language acquisition. In S. Howell, S. Fish, & T. Keith-Lucas, (Eds.), *Proceedings of the 24th annual Boston University conference on language development*, Somerville, MA. Cascadilla Press.

Paribakht, T., & Wesche, M. (1999). Reading and 'incidental' L2 vocabulary acquisition: An introspective study of lexical referencing. *Studies in Second Language Acquisition*, *21*(1), 195-224. https://doi.org/10.1017/S027226319900203X

Pilán, I., Alfter, D., & Volodina, E. (2016). Coursebook texts as a helping hand for classifying linguistic complexity in language learners' writings. In *Proceedings of the workshop on computational linguistics for linguistic complexity,* (pp. 120-126). Osaka, Japan. https://www.aclweb.org/anthology/W16-4114.pdf

Pilán, I., Volodina, E., & Zesch, T (2016). Predicting proficiency levels in learner writings by transferring a linguistic complexity model from expert-written coursebooks. In *Proceedings of the 26th international conference on computational linguistics: Technical papers* (pp. 2101-2111). Osaka, Japan. https://www.aclweb.org/anthology/C16-1198.pdf

Pinker, S. (1994). *The language instinct*. William Morrow and Company.

Pinker, S. (1996). *Language learnability and language development*. Harvard University Press.

Pollard, C. J., & Sag, I. A. (1994). *Head-driven phrase structure grammar*. University of Chicago Press.

Poornamdariyan, T. (1994). *The Persian lesson for foreign Persian learners (For beginners)*. Institute for Humanities and Cultural Studies Publications.

Postman, L., & Keppel, G. (1969). *Verbal learning and memory*. Penguin Books.

Prabhu, N. S. (1987). *Second language pedagogy*. Oxford University Press.

Robinson, P. (2001). Task complexity, cognitive load, and syllabus design. In P. Robinson, (Ed.), *Cognition and second language instruction* (pp. 287-318). Cambridge University Press.

Saffar Moghaddam, A. (2003). *General Persian: Basic constructions*. Council of Extending Persian Language and Linguistics at the Institute for Humanities and Cultural Studies.

Saffar Moghaddam, A. (2008a). *The Persian language* (Vol. 1). Council of Extending Persian Language and Linguistics at the Institute for Humanities and Cultural Studies.

Saffar Moghaddam, A. (2008b). *The Persian language* (Vol. 2). Council of Extending Persian Language and Linguistics at the Institute for Humanities and Cultural Studies.

Saffar Moghaddam, A. (2008c). *The Persian language*. (Vol. 3). Council of Extending Persian Language and Linguistics at the Institute for Humanities and Cultural Studies.

Saffar Moghaddam, A. (2008d). *The Persian language*. (Vol. 4). Council of Extending Persian Language and Linguistics at the Institute for Humanities and Cultural Studies.

Saffran, J., Aslin, R., & Newport, E. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926-1928. https://doi.org/10.1126/science.274.5294.1926

Salton, G. M., Andrew W., & Chung-Shu Y. (1975). A vector space model for automatic indexing. *Communications of the ACM*, *18*(11), 613-620. https://doi.org/10.1145/361219.361220

Samareh, Y. (1989). *Teaching the Persian language* (Vol. 1). Alhoda International Publications.

Samareh, Y. (2005a). *Teaching the Persian language* (Vol. 2). Alhoda International Publications.

Samareh, Y. (2005b). *Teaching the Persian language*. (Vol. 3). Alhoda International Publications.

Samareh, Y. (2005c). *Teaching the Persian language*. (Vol. 4). Alhoda International Publications.

Sarvi, A., Talebnia, G., Pourzamani, Z., & Jahanshad, A. (2019). Assessment readability and understandability of accounting standards by accountants and auditors using Flesch and Cloze Indexes. *Applied Research in Financial Reporting***,** *7*(2), 241-274. http://www.arfr.ir/article_85308_8ee110e57414180e4fc5eec833f18000.pdf?lang=en

Shekari, A., & Najareyan, Z. (2012). A study of the readability of Hedyehaye Asemani textbooks for grades four and five based on Gunning scale. *Research in Curriculum Planning*, *9*(6), 71-79. http://jsr-e.khuisf.ac.ir/article_534233_1f574dc5383e52c94da235658f255a2a.pdf

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford University Press.

Sulistyowati, T. (2019). Bottom-up and top-down listening progress within cognitive constructivist learning theory. *Prominent Journal of English Studies*, *2*(1), 92-100. https://doi.org/10.24176/pro.v2i1.2962

ten Bosch, L., Boves, L., Van Hamme, H., & Moore, R. K. (2009). A computational model of language acquisition: The emergence of words. *Fundamenta Informaticae, 90*(3), 229-249. https://doi.org/10.3233/FI-2009-0016

Thomas, M., & Reinders, H. (2010). Deconstructing tasks and technology. In M. Thomas & H. Reinders, (Eds.) *Task-based language learning and teaching with technology* (pp. 1-13). Continuum International Publishing Group.

Tomasello, M. (1992). *First verbs: A case study of early grammatical development*. Cambridge University Press.

Tomasello, M. (2000). The item-based nature of children's early syntactic development. *Early language development*, *4*(4), 156-163. https://doi.org/10.1016/S1364-6613(00)01462-5

Tomasello, M. (2006). Acquiring linguistic constructions. In D. Kuhn & R. Siegler (Eds.) *Handbook of child psychology* (pp. 255-298). Wiley. https://doi.org/10.1002/9780470147658.chpsy0206

Uchihara, T., Webb, S., & Yanagisawa, A. (2019). The effects of repetition on incidental vocabulary learning: A meta-analysis of correlational studies. *Language learning*, *69*(3), 559-599. https://doi.org/10.1111/lang.12343

Vaezi, M. R., Kokabi, M., & Ebrahimi, M. (2016). Investigation of the readability level of authored fiction books, selected by Children's Book Council of Iran. *Research on Information Science & Public Libraries*, *21*(4), 629-649. http://publij.ir/article-1-1085-fa.pdf

van Rijsbergen, C. J. (1979). *Information retrieval*, 2nd ed. Butterworth-Heinemann.

Widdowson, H. G. (1978). *Teaching language as communication*. Oxford University Press.

Yang, Y., Yu, W., & Lim, H. (2016). Predicting second language proficiency level using linguistic cognitive task and machine learning techniques. *Wireless Pers Commun, 86*, 271-285. https://doi.org/10.1007/s11277-015-3062-2

Zarea Gavgani V., Mirzadeh-Qasabeh, S., Hanaee, J., & Hamishehkar, H. (2018). Calculating reading ease score of patient package inserts in Iran. *Drug Healthc Patient Safety*, *19*(10), 9-19. https://doi.org/10.2147/DHPS.S150428

Zarghamiyan, M. (1998). *Series of teaching the Persian language (From Beginner to Advanced)* (Vol. 1). Council of Extending Persian Language and Linguistics.

Zarghamiyan, M. (2001a). *Series of teaching the Persian language (From Beginner to Advanced)* (Vol. 2). Council of Extending Persian Language and Linguistics.

Zarghamiyan, M. (2001b). *Series of teaching the Persian language (From Beginner to Advanced)* (Vol. 3). Council of Extending Persian Language and Linguistics.

Zeinali, V., Haghparast, A., Damerchilou, M., & Vazifehshenas, N. (2019). Quality and

readability of online health information produced by the Ministry of Health and Medical Education of Iran. *Journal of Health Administration*, *21*(74), 65-74. http://jha.iums.ac.ir/article-1-2798-en.pdf