Exploiting Morphological, Grammatical, and Semantic Correlates for Improved Text Difficulty Assessment

Elizabeth Salesky, Wade Shen †
MIT Lincoln Laboratory Human Language Technology Group, 244 Wood Street, Lexington MA 02420, USA
{elizabeth.salesky, swade}@ll.mit.edu

Abstract

We present a low-resource, language-independent system for text difficulty assessment. We replicate and improve upon a baseline by Shen et al. (2013) on the Interagency Language Roundtable (ILR) scale. Our work demonstrates that the addition of morphological, information theoretic, and language modeling features to a traditional readability baseline greatly benefits our performance. We use the Margin-Infused Relaxed Algorithm and Support Vector Machines for experiments on Arabic, Dari, English, and Pashto, and provide a detailed analysis of our results.

1 Introduction

While there is a growing breadth of reading materials available in various languages, finding pertinent documents at suitable reading levels remains difficult. Information retrieval methods can find resources with desired vocabulary, but educators still need to filter these to find appropriate difficulty levels. This task is often more challenging than manually adapting the documents themselves. Reading level assessment systems can be used to automatically find documents at specific Interagency Language Roundtable (ILR) levels, aiding both instructors and learners by providing proficiency-tailored materials.

While interest in readability assessment has been gaining momentum in many languages, the majority of previous work is language-specific. Shen et al. (2013) introduced a baseline for language-independent text difficulty assessment, based on the ILR proficiency scale. In this work, we replicate and extend their results.

† This work is sponsored by the Defense Language Institute under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

The ILR scale is the standard language proficiency measure for the U.S. federal government. It ranges from no proficiency to native proficiency on a scale of 0-5, with half-level denotations where proficiency meets some but not all of the criteria for the next level (Interagency Language Roundtable, 2013). For second language learners, it is sufficient to use up to ILR level 4. Since proficiency is a continuous spectrum, text difficulty assessment is often treated as a regression problem, as we do here. Though the ILR levels may appear to be discrete categories, documents can fall between levels. The degree to which they do is important for us to measure.

Level	Description
1	***Elementary***: can fulfill basic needs, limited to fundamental vocabulary
2	***Limited working***: routine social demands, gist of non-technical works, elementary grasp of grammar
3	***General professional***: general vocabulary, good control of grammar, errors do not interfere with understanding
4	***Advanced professional***: fluent language use on all levels, only rare & minute errors

Table 1: Description of proficiency at ILR levels

The ILR scale addresses semantic and grammatical capabilities, and to model it appropriately, a system needs to reflect both. The baseline system developed by Shen et al. (2013) uses both term frequency log-weighted (TFLOG) word-usage features and z-normalized word, sentence, and document length features. However, their results are not equally significant across its set of test languages, which this paper addresses with additional features.

The utilization of types for TFLOG weighted vectors is not as representative for morphologically rich languages, where multiple types can represent different word-forms within a single

Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications , pages 155–162,
Baltimore, Maryland USA, June 26, 2014.

paradigm. By incorporating morphology, we can improve our TFLOG vectors' representation of semantic complexity for these languages. We employ the Morfessor Categories-MAP algorithm for segmentation (Creutz & Lagus, 2007). Relative entropy and statistical language models (LMs) can also measure semantic complexity, and class-based language models (cLMs) can give us a measure of the grammatical complexity of the text. All of these methods are low-resource and unsupervised; they can be easily applied to new languages. We have compared their performance to language-specific methods where possible.

The remainder of this paper is structured as follows; Section 2 summarizes previous research on readability assessment. Section 3 introduces our corpus and approach, while Section 4 details our results and their analyses. Section 5 provides a summary and description of future work.

2 Background & Related Work

Early work on readability assessment approximated grammatical and lexical complexity using shallow features like sentence length and the number of syllables in a word, like the prominent Flesch-Kincaid measure, in large part due to their low computational cost (Kincaid et al., 1975). Such features over-generalize what makes a text difficult; it is not always the case that longer words and sentences are more grammatically complex than their shorter counterparts. Subsequent work such as the Dale-Chall model (Dale & Chall, 1995) added representation on static word lists: in this case, one of 3,000 words familiar to 4th graders. Such lists, however, are not readily available for many difficulty scales and languages.

Ensuing approaches have employed more sophisticated methods, such as word frequency estimates to measure lexical complexity (Stenner, 1996) and statistical language models to measure semantic and syntactic complexity, and have seen significant performance gains over previous work (Collins-Thompson & Callan, 2004; Schwarm & Ostendorf, 2005; Petersen & Ostendorf, 2009). In the case of Heilman et al. (2007), the combination of lexical and grammatical features specifically addressed the order in which vocabulary and grammar are acquired by second language learners, where grasp of grammar often trails other markers of proficiency.

The extension of readability research to languages beyond English necessitated the introduction of new features such as morphology, which have long been proven useful in other areas. Dell'Orletta et al. (2011) developed a two-class readability model for Italian based on its verbal morphology. François and Fairon (2012) built a six-class readability model, but for adult learners of French, utilizing verb tense and mood-based features. Most recently, Hancke et al. (2012) built a two-class German reading level assessment system heavily utilizing morphology. In addition to traditional syntactic, lexical, and language modeling features used in English readability research, Hancke et al. (2012) tested a broad range of features based on German inflectional and derivational morphology. While all of these systems were very effective, they required many language-specific resources, including part-of-speech tags.

Recent experiments have several noteworthy characteristics in common. While some systems discriminate between multiple grade-level categories, most are two- or three-class classification tasks between 'easy' and 'difficult' which do not require such fine-grained feature discrimination. Outside of English, there are few multi-level graded datasets; for those that do exist, they are very small, averaging less than a hundred labeled documents per level. Further, though recent work has been increasingly motivated by second language learners, most systems have only been implemented for a single language (Schwarm & Ostendorf, 2005; Petersen & Ostendorf, 2009); Vajjala & Meurers, 2012). The language-specific morphological and syntactic features used by many systems outside of English would make it difficult to apply them to other languages. Shen et al. (2013) address this problem by using language-independent features and testing their work on four languages. In this work, we extend their system in order to improve upon their results.

3 Approach

3.1 Corpus

We conducted our experiments on the corpus used by Shen et al. (2013). The dataset was collected by the Defense Language Institute Foreign Language Center (DLIFLC) for instructional use. It comprises approximately 1390 documents for each of Arabic, Dari, English, and Pashto. The documents are evenly distributed across seven test ILR levels: {1, 1+, 2, 2+, 3, 3+, 4}. This equates to close to

200 documents per level per language. We use an 80/20 train test split.

Lang.	Tokens	Types	Stems	Morphs / Word
Arabic	593,113	84,160	14,591	2.60
Dari	761,412	43,942	13,312	2.61
English	796,406	44,738	35,594	1.80
Pashto	840,673	59,031	20,015	2.34

Table 2: Corpus statistics

The documents were chosen by language instructors as representative of a particular level and range from news articles to excerpts from philosophy to craigslist postings. Three graders hand-leveled each document. The corpus is annotate only with the aggregate scores; we use only this score for comparison. The creation of the corpus took 70 hours per language on average. We assume the ILR scale is linear and measure performance by mean squared error (MSE), typical for regression. MSE reflects the variance and bias of our predictions, and is therefore a good measure of performance uniformity within levels.

3.2 Experimental Design

We compare our results to the best performing Suport Vector Machine (SVM) and Margin-Infused Relaxed Algorithm (MIRA) baselines from Shen et al. (2013). Both of these baselines have the same features: TFLOG weighted word vectors, average sentence length by document, average word length by document, and document word count. We used an implementation of the MIRA algorithm for regression (Crammer & Singer, 2003). We embedded Morfessor for unsupervised morphological segmentation and preprocessed our data as required by this algorithm (Creutz & Lagus, 2007). To verify our results across classifiers, we compare with SVM (Chang & Lin, 2001). We also compare Morfessor to ParaMor (Monson 2009), an unsupervised system with a different level of segmentation aggression, as well as to language-specific analyzers.

Our experiments apply word-usage features, shallow length features, and language models. For the first, we compare TFLOG vectors based on word types, all morphemes, and stems only. For the second, we tested the three baseline shallow length features (average word length in characters per document, average sentence length per document, and document word count) as well as measures of relative entropy, average stem fertility, average morphemes per word, and the ratio of types to tokens. Of these, only relative entropy positively impacted performance, and only its results are reported in this paper. All length features were z-normalized. We compare both word- and class-based language models. We trained LMs for each ILR level and used the document perplexity measured against each as features.

Optimal settings were determined by sweeping algorithm parameters, and Morfessor's perplexity threshold for each language. We conducted a feature analysis for all combinations of word, length, and LM features across all four languages.

4 Results & Analysis

We first replicate the baseline results of Shen et al. (2013) using both the MIRA and SVM algorithms. We find there is very overall little performance difference between the two algorithms, and the difference is language-dependent. It is inconclusive which algorithm performs best.

Algorithm	AR	DA	EN	PA
MIRA	0.216	0.296	0.154	0.348
SVM	0.198	0.301	0.147	0.391

Table 3: Baseline results in MSE, SVM vs. MIRA

Table 3 shows the average MSE across the seven ILR levels for each language. Figure 1 depicts MSE performance on each individual ILR level.

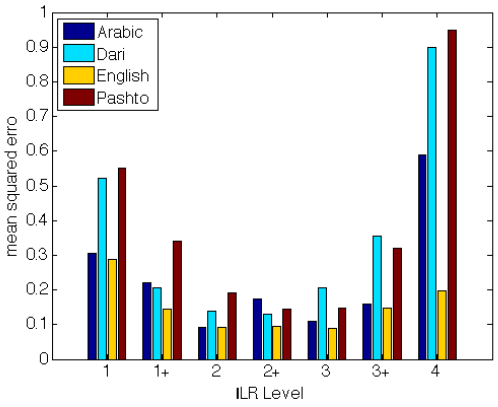

Figure 1: MSE by ILR level, baseline

4.1 Morphological Analysis

Reading level assessment in English does not necessitate the use of morphological features, and so

they have not been researched for this task until recently. Morphology has long been shown to be useful in other areas; it is unsurprising that segmentation should help with this task for morphologically rich languages. What we demonstrate is that unsupervised methods perform similarly to language-specific methods, at a lower cost.

Language	TYPES	MORPHS	STEMS
Arabic	0.216	0.198	0.208
Dari	0.296	0.304	0.294
English	0.154	0.151	0.151
Pashto	0.348	0.303	0.293

Table 4: Average MSE results comparing the use of types, all morphs, and stems for TFLOG vectors. Morfessor algorithm used for segmentation.

Table 4 compares the performance of the baseline, which utilizes types for its TFLOG weighted word vectors, to our configurations that alternatively use all morphemes or stems only. We see that morphological information improves performance for all cases but one, all morphs for Dari, and that using stems only shows the greatest improvement.

Our greatest improvement was seen in Pashto, which has the most unique stems in our dataset both outright and compared to types (see Table 4). Without stemming, TFLOG word vectors were heavily biased by the frequency of alternate word forms within a paradigm. With stemming, which reduced overall MSE compared to the baseline by 16%, the number of word vectors in the optimized configuration increased by 18%, and were much more diverse, reflecting the actual semantic complexity of the documents. We posit that the reason Dari, which has a similar ratio of morphemes per word to Pashto, does not improve in this way is due to its much smaller and more uniform vocabulary in our data. Our Pashto documents have 1.5 times as many unique words as our Dari, and in fact, with stemming, the number of word vectors utilized in our optimized configuration was reduced by 20%, as fewer units were necessary to reflect the same content.

We compare our results using Morfessor to another unsupervised segmentation system, ParaMor (Monson 2009). ParaMor is built on a different mathematical framework than Morfessor, and so has a very different splitting pattern. Morfessor has a tunable perplexity threshold that dictates how aggressively the algorithm segments. Even set at its highest, ParaMor still segments much more aggressively, sometimes isolating single characters, which can be useful for downstream applications (Kurimo et al. 2009). This is not the case here, as shown in Table 5. All further results use Morfessor for stemming.

Algorithm	AR	DA	EN	PA
Morfessor	0.208	0.294	0.151	0.293
ParaMor	0.227	0.321	0.158	0.301

Table 5: Comparison of unsupervised segmenters

To our knowledge, no Pashto-specific morphological analyzer yet exists for comparison. However, in lacking both a standardized writing system and spelling conventions, one word in Pashto may be written in many different ways (Kathol, 2005). To account for this, we normalized the data using the Levenshtein distance between types. We swept possible cutoff thresholds up to 0.25, evaluated by the overall MSE of the subsequent results. Using normalized data did not improve results; in many cases the edit distance between alternate misspellings is just as high or higher as the distance between word types.

We believe that the limited change in Dari performance is primarily related to corpus characteristics; relatively uniform data provides low perplexity, making it more difficult for Morfessor to discover all morphological segmentations. Using the Perstem stemmer in place of Morfessor, the number of word vectors in the optimized system rose 143% and our results improved 8%. This increase affirms that Morfessor is under-splitting. Perstem is tailored to Farsi, and while the two dialects are mutually intelligible, they have grammatical, phonological, and loan word differences (Shah et al. 2007).

We highlight that the overall MSE of all configurations in Table 4 vary only 2% for English, with identical results using all morphs and only stems. This is expected, as English is not morphologically complex. Given the readily available rule-based systems for English, we compared results with Morfessor to the traditional Porter and Paice stemmers, as well as the multi-lingual FreeLing stemmer, as seen in Table 6.

Performance variance between all analyzers of only 3% points us to the similar and limited grammatical rules found in the different algorithms, as well as the relatively limited number of unique

Baseline	Morf.	Porter	Paice	FreeLing
0.154	0.151	0.149	0.148	0.153

Table 6: Comparison of English segmenters

stems and affixes to be found in English. Topical similarities in our data are also possible.

Like Pashto, Arabic has a rich morphological structure, but in addition to affixes it contains templatic morphology. It is difficult for unsupervised analyzers not specifically tailored to templatic morphology to capture non-contiguous morphemes. Here, Morfessor consistently segments vowelized types into sequences of two character stems. When compared with MADA, a rule-based Arabic analyzer (Habash, 2010), we found that Morfessor outperformed MADA by 10%. This is likely because the representations present in the dataset are what is significant; if a form is 'morphologically correct' but perpetuates a sparsity problem, linguistically-accurate stemming will not help. Neither stemmer contributes much to Arabic results, however, as MIRA does not weight word-usage features very heavily for either Arabic analyzer.

4.2 Relative Entropy and Word LMs

As mentioned in Section 2, traditional features like document word count and average sentence length overstate the importance of length to difficulty. To capture the significance of the length of the document, rather than merely the length itself, we utilized relative entropy. Relative entropy, also known as the Kullback-Leibler divergence (KL), is a measure of the information lost by using one probability distribution as compared to another. Expressed as an equation, we have:

$$D(p,q) = \sum_{x \in \varepsilon} p(x) \log \frac{\mathrm{p}(x)}{\mathrm{q}(x)}. \qquad (1)$$

In this work, we are comparing a unigram probability distribution of a document $q(x)$ to a uniform distribution over the same length $p(x)$. This provides both a measure of the semantic and structural complexity of a document, allowing us to differentiate between documents of similar length. Figure 2 shows the normalized distribution of the relative entropy feature for Pashto.

The separability of ILR levels suggests we will be able to discriminate between them. As demonstrated by the improved performance in Figure 3, where the inclusion of relative entropy is super-

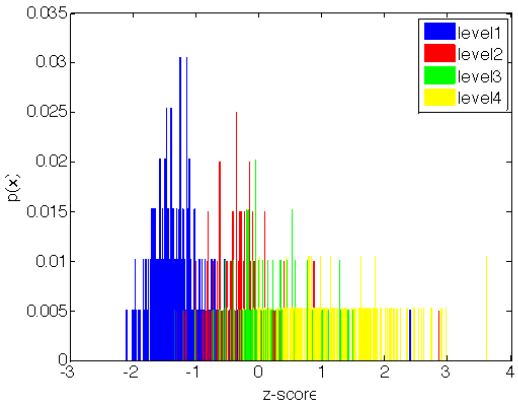

Figure 2: Pashto, normalized KL distribution

imposed over the baseline, this feature greatly contributes to the separability of outlier levels of our corpus. Common z-scores between levels 2 and 3 explain the system's poorer performance on the ILR levels 2.0 and 2.5 (Figure 3). Adding the relative entropy feature to the baseline produced an average MSE reduction of 15%.

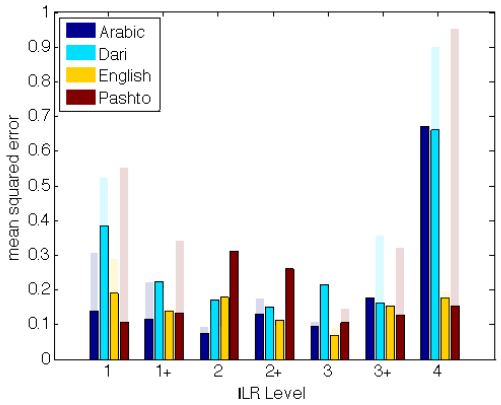

Figure 3: MSE by ILR level, baseline +stems +KL

The combination of stemming for TFLOG vectors and relative entropy together is more effective than either alone. Further removing document word count improved performance by an average 1%. As seen in Figure 3, the combination of all these changes produces significant gains over the baseline, particularly in Dari and Pashto. The combination configuration reduced overall MSE by 52% for Pashto documents and by 18% for Dari. From Figure 3 above, we see that the +stems+KL configuration exhibits very poor performance in Arabic level 4, and on outlying levels for Dari. While these MSE values are clear outliers in this figure, they values are less than 0.1 greater than their MIRA baseline coun-

terparts. This may be due to data similarity between level 3+ and 4 documents, or MIRA may have been overfit during training. In contrast, the variance for English and Pashto is much smaller; overall, the variance has been greatly reduced.

Statistical language models (LMs) are a probability distribution over text. An n-gram language predicts a word w_n given the preceding context $w_1 ... w_{n-1}$. We used the SRI Language Modeling Toolkit to train LMs on our training data for each ILR level (Stolcke, 2002). To account for unseen n-grams, we used Kneser-Ney smoothing. To score documents against these LMs, we calculate their perplexity (PP), a measure of how well a probability distribution represents data. Perplexity represents the average number of bits necessary to encode each word. For each document in our dataset, we use the perplexities against each ILR level LM as features in MIRA. We compared n-gram orders 2-5, and while we found an average decrease of 3% MSE between orders 2 and 3 across languages, there was a difference of less than 1% between 3-gram and 5-gram LMs.

Features	AR	DA	EN	PA
baseline	0.216	0.296	0.154	0.348
+stems +KL	0.208	0.269	0.147	0.173
+LM	0.208	0.176	0.117	0.171
+LM -WVs	0.567	0.314	0.338	0.355
+stems +KL +LM	0.168	0.167	0.096	0.137

Table 7: Average MSE results comparing features from Sections 4.1 and 4.2. LMs are order 5.

As we can see from Table 7, the addition of language models alone can provide a huge measure of improvement from the baseline. For Arabic and Pashto, it is the same improvement seen by stemming TFLOG vectors and adding relative entropy. For Dari and English, however, the performance improvement is unmatched by any other features presented thus far. We compare these results to the same configuration without TFLOG vectors, in order to measure the overlap between these features; see Table 7. Based on the relative results, it seems that word vector and LM features are orthogonal. The addition of all three new features (stemmed word vectors, relative entropy, and language models) provides considerable further improvement upon any previous configuration. It appears that the interactions between these features have a further positive influence on our discriminative ability.

4.3 Class-Based LMs

It is possible to group words based on similar meaning and syntactic function. It is reasonable to think that the probability distributions of words in such groups would be similar (though not the same). By assigning classes to words, we can calculate the probability of a word based not on the sequence of preceding *words*, but rather, *word classes*. Doing so decreases the size of resulting models and also allows for better predictions of unseen word sequences. Sparsity is a concern with language models, where we rely on the frequency of sequences, not just words. Using word classes assuages some of this concern. These word classes are generated in an unsupervised manner. We train our class-based language models (cLMs) using c-discounting to account for data sparsity.

Features	AR	DA	EN	PA
baseline	0.216	0.296	0.154	0.348
+LM	0.208	0.176	0.117	0.171
+cLM	0.130	0.286	0.144	0.211
+LM +cLM	0.094	0.155	0.051	0.084
+stems +KL +LM +cLM	0.092	0.152	0.049	0.079

Table 8: Average MSE results comparing all features. LMs and cLMs are order 5.

Class-based and word-based LMs each help different languages in our test set. The two types of LMs model different information, with word-based LMs providing a measure of semantic complexity and class-based modeling grammatical complexity. As seen in Table 8, the combination of this complementary information is highly beneficial and strongly correlated to ILR level. We see average MSE reductions of 56%, 48%, 67%, and 77% in Arabic, Dari, English, and Pashto, respectively, using both types of language model.

Algorithm	AR	DA	EN	PA
MIRA	0.091	0.156	0.049	0.079
SVM	0.089	0.159	0.069	0.070

Table 9: Final system results, comparing avg. MSE with the MIRA and SVM algorithms

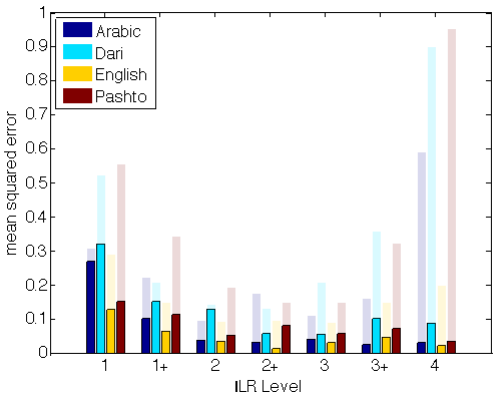

Figure 4: Comparison of final configuration with all features to baseline by MSE, MIRA algorithm

The further inclusion of TFLOG stemming and relative entropy reduces average MSE an additional 1%. Figure 4 reflects this configuration's performance across the seven ILR levels.

Figure 4 superimposes our final error results over those of the baseline. It is clear that error has become much less language-specific; performance on all seven ILR levels has become considerably more consistent across the four languages, as has the accuracy at each individual ILR level. It seems likely that our error measures would be similar to inner-annotator disagreement, a measure that we would like to quantify in the future.

We find that our results are significant across classifiers. Table 9 shows the performance of our final feature set with both MIRA and SVM. The MSE exhibits the same trends across ILR levels and languages with both algorithms. The average difference in error between the algorithms remains the same as it was with the baseline features.

5 Conclusions and Future Work

Our experiments demonstrate that language-independent methods can improve text difficulty assessment performance on the ILR scale for four languages. Morphological segmentation for TFLOG word vectors improves our measure of semantic complexity and allows us to do topic analysis better. Unsupervised methods perform similarly to language-specific and linguistically-accurate analyzers on this task; we are not sacrificing performance for a language-independent system. Relative entropy gives structural context to more traditional shallow length features, and with word-based LM features provide another way to measure semantic complexity. Class-based LM features measure grammatical complexity and to some degree account for data sparsity issues. All of these features are low-cost and require no language-specific resources to be applied to new languages. The combination of all these features significantly improves our performance as measured by mean square error across a diverse set of languages.

We would like to expand our work to more diverse languages and datasets in future work. There is room to improve upon features described in this paper, such as new frequency-based measures for word vectors and unsupervised morphological segmentation methods. In the future, we would like to directly compare inner-annotator error and well-known formulas with our results. It would also be interesting to look at performance on subsets of the corpus to test dependence on dataset size. We would also like to investigate the ILR scale; while we assume that it is linear, this is not likely to be the case.

Acknowledgments

This paper benefited from valuable discussion with Jennifer Williams.

References

J. Chall, E. Dale. 1995. Readability revisited: The new Dale-Chall readability formula. *Brookline Books*, Cambridge, MA.

C-C. Chang, C-J. Lin. 2001. LIBSVM: a library for support vector machines. *Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm*.

K. Collins-Thompson, J. Callan. 2005. Predicting reading difficulty with statistical language models. *Journal of the American Society for Information Science and Technology* 56(13), 1448-1462.

K. Crammer, Y. Singer. 2003. Ultraconservative Online Algorithms for Multiclass Problems. *Journal of Machine Learning Research*, 3(2003):951-991.

M. Creutz, K. Lagus. 2007. Unsupervised models for morpheme segmentation and morphology learning. *Association for Computing Machinery Transactions on Speech and Language Processing (ACM TSLP)*, 4(1):1-34.

F. Dell'Orletta, S. Montemagni, G. Venturi. 2011. Read-it: Assessing readability of italian texts with a view to text simplification. *Proceedings of the 2nd Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)* 73-83.

T. François, C. Fairon. 2012. An AI readability formula for French as a foreign language. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing (EMNLP) and Computational Natural Language Learning (CoNLL)*. 466-477.

N. Habash, O. Rambow, R. Roth. 2010. Mada+Tokan: A toolkit for arabic tokenization, diacritization, morphological disambiguation, pos tagging, stemming and lemmatization. *Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR)*.

J. Hancke, S. Vajjala, D. Meurers 2012. Readability Classification for German using lexical, syntactic, and morphological features. *Proceedings of CoLING 2012: Technical Papers*, 1063-1080.

K.S. Hasan, M.A. ur Rahman, V. Ng. 2009. Learning-Based Named Entity Recognition for Morphologically-Rich, Resource-Scarce Languages. *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 354-362.

M. Heilman, K. Collins-Thompson, J. Callan, M. Eskenazi. 2007. Combining Lexical and Grammatical Features to Improve Readability Measures for First and Second Language Texts. *Proceedings of NAACL HLT*, 460-467.

Interagency Language Roundtable. ILR Skill Scale. `http://www.govtilr.org/Skills/ILRscale4.htm`. 2013.

A. Jadidinejad, F. Mahmoudi, J. Dehdari. 2010. Evaluation of perstem: a simple and efficient stemming algorithm for Persian. *Multilingual Information Access Evaluation Text Retrieval Experiments*.

A. Kathol, K. Precoda, D. Vergyri, W. Wang, S. Riehemann. 2005. Speech translation for low-resource languages: The case of pashto. *Proceedings of INTERSPEECH*, 2273-2276.

J.P. Kincaid, R.P Fishburne Jr., R.L. Rodgers, and B.S. Chisson 1975. Derivation of new readability formulas for Navy enlisted personnel. *Research Branch Report, U.S. Naval Air Station, Memphis*, 8-75.

M. Kurimo, V. Turunen, M. Varjokallio. 2009. Overview of Morpho Challenge 2008. *Evaluating Systems for Multilingual and Multimodal Information Access*, Springer Berlin Heidelberg, 951-966.

C. Monson. 2009. ParaMor: From Paradigm Structure to Natural Language Morphology Induction. *PhD thesis. Carnegie Mellon University*.

R. Munro, C.D. Manning. 2010. Subword Variation in Text Message Classification. *The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 510-518.

M. Padr. 2004. FreeLing: An Open-Source Suite of Language Analyzers. *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*.

C.D. Paice. 1990. Another Stemmer. *SIGIR Forum*, 24:56-61.

S. E. Petersen and M. Ostendorf. 2009. A machine learning approach to reading level assessment. *Computer Speech and Language*, 23(2009):89-106.

M.F. Porter. 1980. An algorithm for suffix stripping. *Program*, 14(3): 130-137.

A. Ratnaparkhi. 1997. A simple introduction to maximum entropy models for natural language processing. *IRCS Technical Reports Series*, 81.

S. E. Schwarm and M. Ostendorf. 2005. Reading Level Assessment Using Support Vector Machines and Statistical Language Models. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*.

M.I. Shah, J. Sadri, C.Y. Suen, N. Nobile. 2007. A New Multipurpose Comprehensive Database for Handwritten Dari Recognition. *11th International Conference on Frontiers in Handwriting Recognition*, Montreal, 635-40.

W. Shen, J. Williams, T. Marius, E. Salesky. 2013. A language-independent approach to automatic text difficulty assessment for second-language learners. *Proceedings of the 2nd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR) 2013*.

A. Stolcke. 2002. SRILM - an extensible language modeling toolkit. *Proceedings of the ICSLP*, vol. 2, 901-4.

S. Vajjala, D. Meurers. 2012. On improving the accuracy of readability classification using insights from second language acquisition. *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP. Association for Computational Linguistics, 2012*. 163-173.