# A Novel Use of MT in the Development of a Text Level Analytic for Language Learning

**Carol Van Ess-Dykema** cvanessdykema@casl.umd.edu
U.S. Department of Defense, 9800 Savage Road, Suite 6908, Ft. Meade, MD 20755-6908
**Salim Roukos** roukos@us.ibm.com
IBM T.J. Watson Research Center, Yorktown Heights, NY 105989
**Amy Weinberg** aweinberg@casl.umd.edu
University of Maryland, Center for Advanced Study of Language, 7002 52nd Avenue, College Park, MD 20742

**Abstract**

In the current budget climate, it is important to investigate multiple uses for technologies that have benefitted from U.S. Government investment. Authentic materials indexed at appropriate learning levels are a requirement of several foreign language training activities. In this paper, we propose an approach that automatically annotates texts with language proficiency/ difficulty levels. Our approach is novel in that it uses independently available machine translation output of the source language into English coupled with an English-trained automatic text leveling analytic. This approach precludes text leveling annotation for each new language, though it requires a machine translation system. We report our initial results for Farsi document leveling. This automatic system is introduced as part of a wider adaptive platform currently under development called LanguageNation.

## 1 Introduction

The LanguageNation Human Language Technology (HLT)-enabled adaptive language learning platform will provide anytime, anywhere, any-device language learning. It offers several important advancements that are currently unobtainable: HLT-based content analysis to support automated adaptivity in language learning and fetching of appropriate, authentic language materials "on the fly, and learner models indicating subject matter interest, language proficiency, and individual cognitive style. This paper concentrates on the uses of machine translation (MT) to support acquisition of authentic foreign language materials that are appropriately calibrated by the human proficiency level needed to process them, indexed according to the U.S. Government (USG) Interagency Language Roundtable (ILR) proficiency scale.

### 1.1 LanguageNation – an adaptive learning platform developed by a U.S. Government, industry, and academe partnership

LanguageNation represents the most innovative platform currently available in foreign language instruction. The platform will collect data on a learner's progress and adapt instructional content to the particular learner's needs, abilities, and skills. Each iteration of development moves the platform closer to this goal. In traditional language instruction, decisions about instruction content are made by curriculum developers, instructors and, more

rarely, learners themselves. In the case of adaptive learning, the platform is trained to make these decisions so that the learner is provided with optimal input and is invited to engage with the language features that are most appropriate for the learner at a given time. That is, the LanguageNation platform will enable self-directed learners to access anytime, anywhere language instruction that is tailored to their specific learning profile. Because this adaptivity enables tailored learning with little to no involvement of a human instructor (or coach), the LanguageNation platform has the potential to support learners at all levels of proficiency regardless of their physical proximity to a language instructor, while also empowering language instructors to tailor instruction to the individual. The LanguageNation platform will integrate a number of HLT tools that will support the automatic fetching of relevant, authentic language materials from the Web in support of effective language learning. Users' goals and needs (e.g., topics of interest, desired outcomes of learning) will inform the selection and presentation of authentic materials from a range of media, audio clips, videos, and blogs. The use of authentic materials is important for a learner to develop advanced proficiency in the target language and enhances the relevance of the learning materials to the students' needs. This increases "stickiness" – that is, the degree to which materials promote a high level of engagement and commitment to the training. Moreover, by automating the fetching of authentic materials, instructors will no longer have the burden of finding relevant, timely materials for students, thereby freeing up their time to focus on the students.

HLT assets in LanguageNation include a wide variety of analytics that are designed to automatically process and understand human-generated content. Traditional HLT investment has focused on development of analytics to create annotations to describe the content of documents considered one at a time. More recently, widespread Governmental, academic and industrial investment has encouraged the development of analytics that can provide deeper understanding of content spanning multiple documents, media and languages. These efforts are being leveraged to create advanced information retrieval and dialogue management opportunities and modified for LanguageNation to capture more authentic language learning materials tailored to promote advanced proficiency.

The LanguageNation data processing and management framework provides the necessary infrastructure to create bindings between learners, analytics, and data. Traditional solutions emphasized monolithic data processing instances that were local to both the learner and the data. Following more recent trends that are pursuing distributed processing and storage across multiple platforms in support of geographically dispersed users, LanguageNation will create anywhere, anytime, any-device language learning opportunities.

LanguageNation is a unique partnership among the USG, industry and academe; specifically, the Department of Defense, IBM, and the University of Maryland Center for Advanced Study of Language (CASL). LanguageNation's success is dependent upon the contributions from each of the three entities.

### 1.2 Text Leveling for a USG Language Proficiency and Language Testing Framework

The Interagency Language Roundtable (ILR) was established to coordinate and share information concerning USG language-related activities. The ILR scale is a set of descriptions of abilities to communicate in a language. The scale [1] of 11 levels ranges from level 0 to level 5 using half steps (0, 0+, 1, 1+, 2, 2+, 3, 3+, 4, 4+, 5). Level 0 indicates no proficiency, level 5 indicates functional native proficiency and level 3 indicates general professional proficiency.

The ILR Guidelines are used by the USG to rate a USG professional's language proficiency. Proficiency as measured using the ILR scale currently sets criteria for students graduating from the Defense Language Institute Foreign Language Center (DLI FLC) and determines Foreign Language Incentive Payment eligibility for many USG professionals. There has been significant Government investment in designing assessment instruments to rate proficiency at specific ILR levels of test-takers. Slightly more controversially, this effort has extended to determining the proficiency level of the language used in test items as well as in the materials from which test items are constructed.[1] Regardless of differing views on this topic, it is clear that some notion of proficiency must be part of the process of selecting either materials or material/activity pairs for test development. One would not ask a beginning language learner, for example, to examine a philosophical tract for its intended meaning as this would normally require language knowledge above the learner's current language proficiency.

## 2 Definition of the Problem

A standard requirement for USG testing instruments is the use of authentic materials. Additionally, USG instructors routinely provide their students with authentic materials as part of their normal lesson designs. Therefore an automatic proficiency leveling capability that can support either item development for testing or classroom use is desirable to the extent that it can decrease the cost of providing authentic materials for these uses. Instructors/test developers spend significant time manually leveling texts for their proficiency level. In IBM-directed studies, the average annotation time for short passages was between 2.3 and 7.2 minutes for English documents by four annotators (kt, mn, mr, rx) trained in a 10 hour ILR annotation course as we see in Table 1 (Roukos, Quin, & Ward, 2014).

| | kt | mn | mr | rx |
|---|---|---|---|---|
| avg time | 0:02:31 | 0:07:24 | 0:03:14 | 0:03:30 |
| avg abs error | 0.64 | 0.57 | 0.55 | 0.67 |
| num errors | 34 | 28 | 28 | 36 |
| large err | 3 | 3 | 4 | 7 |

Table 1. Text leveling: human annotator performance.

In order to ensure the validity of proficiency leveling, the normal procedure is to have texts leveled independently by two annotators and adjudicated by a senior annotator, whose judgment is used for final scoring. Annotation thus becomes a laborious, time-intensive and expensive process. Manual annotation for the purposes of the classroom becomes infeasible as instructors often desire to use authentic, recent material. They select documents describing current events to present real world language that a USG professional is likely to encounter and to maintain learner interest based on topicality. Even determinations by a single judge that can take up to seven minutes per passage are prohibitive, since many materials subjected to this process might be rejected as too easy or too difficult, thus multiplying the annotations required.

[1] There is some controversy regarding the issue of whether a proficiency level can be specified with respect to a test or item text alone, or whether the same text's difficulty can vary depending on the assessment activity to be performed on the basis of that text. Agencies differ with respect to whether determining text difficulty level directly is a requirement for test production or not.

Therefore "eyeballing" or other informal proficiency determinations are often used for daily selection. This can lead to a disconnect between the learning materials and materials that the learner may confront at test time.

Beyond time cost, there is some question regarding whether the manual indexing process is actually reliable. In an early CASL study, it asked senior annotators (in fact the professionals who developed the text proficiency annotation standards) to index approximately 50 passages for their proficiency level ranging from level 1 to level 4. Their consistency rating using kappa statistics was only 0.63. While this was a small sample, it suggests a rather low gold standard for reliability (Keenan, Resnik, Skaggs, & Weinberg, 2005).

The previous discussion underlines two potential advantages for automated text leveling:

1. It may decrease development time and resulting labor cost for material selection as well as item development for USG test development.
2. It may increase reliability of the resulting text level determinations.

The question we begin to explore in this paper is whether we can drive down the cost in terms of the time and labor required to produce automated text proficiency leveling systems by leveraging independently available HLT in which the USG is likely to have a sustained investment? We examine the use of MT for this purpose.

## 3 Previous Work

Early work on text proficiency levels addressed the readability of a document based on the Flesch Reading Ease Formula, which uses two simple length features: the average number of words per sentence and the average number of syllables per word. There have been various attempts at exploring weighting these features (using linear regression models) to improve the accuracy of predicting different readability levels (Kincaid, Fishburne, Rogers, & Chissom, 1975). More recent work uses a richer feature set including the following (Petersen & Ostendorf, 2009; Schwarm & Ostendorf, 2005):

- average sentence length
- average number of syllables per word
- Flesch-Kincaid score
- six out-of-vocabulary (OOV) rate scores
- syntactic parse features
- 12 language model perplexity scores

Collaboration between MIT's Lincoln Laboratory and the DLI FLC confirmed that automatic text leveling for use in both learning and testing materials was possible. Shen et al used a corpus of 200 documents for each of seven proficiency levels (1, 1+, 2, 2+, 3, 3+, 4) for a given language (Shen, Williams, Marius, & Salesky, 2013). In their data, each of the texts was labeled by two independent linguists expertly trained in ILR level scoring. The ratings from these two linguists were then adjudicated by a third linguist. They did not provide inter-annotator agreement measures but took the adjudicated decision as the reference truth for both training and testing their system.

Using the fine-grained ILR level training data, Shen et al developed regression models of text proficiency and proposed the use of the mean squared error (mse) metric where the plus-levels were mapped to the mid-point (e.g. 2+ is 2.5). They used an 80/20 split for their training and test data and built a separate regression model for each of four languages.

Their best results were a mean squared error (mse) of 0.2 for Arabic, 0.3 for Dari, 0.15 for English, and 0.36 for Pashto. These would correspond to a root mean squared error (rmse) of 0.45, 0.55, 0.39, and 0.60 for each of the languages, respectively. It is important to note that a change of one level is an interval of 0.5 in their study.

In subsequent experiments to those above they used two types of features: length features and word-usage features. The length features were three z-normalized length features:

- average sentence length (in words) per document
- number of words per document
- average word length (in characters) per document.

The word-usage features were weighted word frequencies using TF-LOG weighted word frequencies on bag-of-words for each document. They compared length-based features which are not lexical to word-usage features which are lexical items. The lexical features reduce the mse by 14% (Dari) to about 80% (Pashto).

The word-usage features, while improving the output scores considerably require some comment. We surmise that the data they used are more homogeneous than what are required in second language acquisition (SLA) and may be influencing the significant performance improvement due to the word-usage features. The problem is that their leading examples of useful lexical features for English (which yielded a reduction of mse by 58%) appear to be topical. For example for level 3, the highest ten lexical features, shown in Table 2, appear to be U.S. politics centric.

| Word | Weight |
|---|---|
| obama | 1.739 |
| to | 1.681 |
| republicans | 1.478 |
| ? | 1.398 |
| than | 1.381 |
| more | 1.365 |
| cells | 1.355 |
| american | 1.338 |
| americans | 1.335 |
| art | 1.315 |

Table 2. Top 10 word-usage features
and their weights for level 3 proficiency.

While it is hard to make solid claims about topicality without having access to their data, we are concerned about the robustness of their results. We expect a requirement for topic change over time and place for SLA content, which could lead to degradation of system performance. For example, what would happen when the news is about French politics? Surely, the personal names and political parties will be different from the highest indicators shown above.

Nevertheless, it is clear that acceptable text proficiency levels can automatically be determined, making document selection by a test developer or instructor faster and less

expensive. These online savings need to be evaluated against the time and labor cost for the development of the underlying computational system on a per language basis, however.

Shen et al's approach requires a training corpus of approximately 200 documents per distinct level. Within the ILR scale, there are seven levels of interest so this requires the efforts of annotators who are skilled both in the language and in the ILR Guidelines. The first problem is that there are often only a small number of professionals available to the USG who can meet these requirements, slowing system development. Assuming that qualified professionals are available and assuming an average of five minutes per annotator + review by an adjudicator, the process of corpus preparation for system training would take approximately two months. While Shen et al do not discuss the inter-annotator agreement, we will assume that this process guarantees reliable text leveling.

We are investigating a solution that does not require the development of large human-annotated text proficiency level training corpora. We are currently investigating the utility of MT for accomplishing this goal.

## 4 Our Solution

The USG has a long history of investment in MT development. Our proposal is to leverage this investment by translating incoming source text into English using MT. Once a document is translated, text leveling will occur on the English output. The advantage of this approach is its cost savings and its feasibility for use on new languages. A text leveling system is built once using a seed set of translations, eliminating the need for text level annotation in multiple languages beyond the development of the initial system. We report on our initial results for automatic Farsi document proficiency leveling.

CASL provided IBM with a document collection with reference truth ILR text leveling annotations provided by expert annotators at the National Foreign Language Center at the University of Maryland. The data came from 54 languages, excluding English, and each document included a human translation of the foreign language document into English. The documents covered five broad topical areas and are evenly split between written texts (4,500) and human transcriptions (5,000). Table 3 shows the division by topic. The leveling conventions used in this training set conflated some of the seven ILR levels (0+/1, 1+/2, 2+/3, 3+/4).

| Culture/ Society | Defense/ Security | Ecology/ Geography | Economics/ Politics | Science/ Technology |
|---|---|---|---|---|
| 3635 | 1046 | 823 | 2904 | 1007 |

Table 3. Text leveling data set.

Building the text leveling system required approximately 2-person months. This included building a training and test corpus and conducting several assessments of the system. It is important to recall that this system is built once and used for all subsequent languages. The first experiment uses a training corpus of 2,000 documents and then uses the English translations of these documents to build a text leveling system trained with the following features:

- number of words per document
- average sentence length in words per document
- average word length in characters per document
- ratio of count of unique words (types) to total words
- pronoun histogram
- POS bigrams
- log term frequency

It measured the performance by the assigned level accuracy, the mean squared error (mse), and its corresponding root mean squared (rms) error. It used a maximum entropy regression model.

These measures can be used to assess the accuracy of the text leveling system, when trained on well- structured human translation. The next experiment, with the larger training set repeated the same process but increased the size of the training set to approximately 9000 documents.

The performance of the system in the smaller experiment was tested using 125 test documents.

When IBM used the first three features, which are similar to the basic length features of earlier work, the level assignment accuracy is 66%, and the mse is 0.37 with an rms of 0.60. Adding the remaining features listed above improves the accuracy to 77% and reduces the rmse to 0.51. Table 4 shows the confusion matrix for the full feature set.

| | | | | |
|---|---|---|---|---|
| 0.75 | - | - | 2 | - |
| 1.75 | - | 15 | 5 | 1 |
| 2.75 | - | 4 | 70 | 1 |
| 3.75 | - | 1 | 15 | 11 |

Table 4. Confusion matrix between
the four levels using the full feature set classifier.

To evaluate the effect of MT on text leveling performance, IBM identified the largest subset of textual material by source language in the smaller (2000 document) batch of data which turned out to be Farsi. It used a Farsi test set of 60 documents. It used a phrase-based Farsi-English translation system to produce the MT version of the documents. Phrase-based translation provides mid-level quality translation and thus serves as a useful choice because phrase-based systems can be easily developed for new languages with reasonably sized training corpora. The text leveling system was run on Farsi machine translated documents.

Table 5 compares human translation to MT in terms of accuracy, mse, and rms error. The results show that MT is relatively close to human translation although the rms on Farsi at 0.64 is higher than on the original set of 125 documents at 0.51.

| | accuracy | mse | rmse |
|---|---|---|---|
| Human translation | 65% | 0.41 | 0.64 |
| Machine translation | 57% | 0.47 | 0.69 |

Table 5. Performance with human and machine translation.

Next, IBM repeated the same procedure, but with a larger training corpus to see whether this would improve system performance.

Table 6 shows the results using the full set of 9,000 documents, which indicates the paucity of data for the first level and the dominance of the third level. The data were provided with a two-level score (1+/2) which IBM averaged to 1.75.

| Level | 0.75 | 1.75 | 2.75 | 3.75 |
|---|---|---|---|---|
| Count | 148 | 2,214 | 5,531 | 1,569 |

Table 6. Count of documents for each of the four proficiency levels.

IBM compared the small and large training and test conditions. As can be seen in Table 7, the model trained on the smaller partition of data had an rms error increase to 0.69 on the large test set. The large training set reduces the rms error from 0.69 to 0.54 on the large test set.

| train/test | small | large |
|---|---|---|
| small | 0.63 | 0.69 |
| large | 0.58 | 0.54 |

Table 7. RMSE using both the large and small training and test sets.

It is difficult to compare the performance of the DLI FLC system and the IBM system due to several factors. The first is that the leveling conventions were different for the gold standard in each case, and the second, is that the languages that were labeled for proficiency level are different. However, some information trends inform future work. The IBM system seems comparable in performance to systems using only length features, increasing rmse by approximately seven percent. We believe that this decrement stems from a) the decision to determine proficiency level from translation, rather than on the original source text, and b) the decision to base proficiency level on MT output. This quality decrement may be offset by the fact that the text proficiency leveling system does not require rebuilding for every new language. The cost then, of text leveling becomes the cost of building the required MT system (which often exists and is available at no cost) plus some integration cost. This protocol may be sufficiently cost-effective to warrant development to identify effective learning materials in multiple languages for daily classroom use. The limitation of this approach is obviously its

dependence upon an available and adequate MT engine. When no MT system exists, as will be the case for many low-resource languages, this approach is not feasible.

An obvious question to ask is what level of accuracy is required for a particular language learning application? For example, selection of passages for applications such as testing material, may require very high accuracy and thus require better results than those afforded by solely our MT-based technique. However, one can ask whether the filtering afforded by an MT-based approach would reduce development time of testing materials or whether automatic filtering would produce authentic proficiency-leveled documents sufficient to underpin development of classroom materials? In the next section we describe an opportunity to conduct this type of assessment.

## 5 Planned Use of Text Level Analytic in USG Schoolhouse

USG schoolhouse instructors will teach a Spanish-Portuguese cross-training course using the LanguageNation platform for the first time in its September 2014 course offering. This includes use of the text level analytic to automatically fetch Portuguese materials from the Web. The platform will use the learner profile containing an individual learner's proficiency goal to fetch materials at the optimal proficiency level for the individual learners. The platform will provide the instructors a proxy ILR level of proficiency needed to process each document retrieved. The instructor will validate or modify the proxy level which will allow us to assess the value of the system. At the close of the cross-training course, the LanguageNation team will perform an analysis of the results of the instructors' validations of the automatic text level analytic. The team will use the results to refine the analytic development.

## 6 Future Work

The team is eager to investigate features that will improve either the output quality of the text leveling system or its application to languages for which MT systems do not exist. To improve the quality of the system output, we are investigating methods to ensure that length features are better preserved when translating from the source language to English. We will also investigate the addition of topic neutral lexical features to better characterize text difficulty.

We are also interested in investigating the use of proxies for MT such as a statistical bilingual phrase lexicon extracted from aligned sentences for languages where we do not have an MT system. Again, we will assess whether the filtering quality of such a system is "good enough" for practical application.

MT may also provide input to adaptive learning platforms beyond text proficiency leveling. For example, the LanguageNation platform incorporates exercises with source language sentences that have automatically been rendered into English and the learner is called upon to correct the MT output. The system produces these activities automatically using sentences containing words in the individual learners' targeted wordlist. In the future, we intend to expand the system beyond the word level to phrasal collocations.

Targeted learning also involves supplementing this exercise type with learner models to pinpoint errors that we predict learners will make. At the word level, this could be the set of non-cognates for Spanish professionals learning Portuguese. MT-based error correction in this type of exercise is particularly effective because many USG language professionals are required to perform translation tasks. These exercises provide language learning and task-relevant

language practice. The team intends to investigate when in the learning cycle MT-based exercises should be introduced.

## 7 Conclusion

Roukos et al built a text leveling system using an English training set of approximately 9,000 documents. The rms error of 0.54 achieved is comparable to the earlier work of Shen et al which had an average rms error across four languages of 0.50. The IBM approach relies upon the availability of MT engines instead of annotating a large set of training data for each new source language. This method shows promise for saving cost and labor both in terms of system development and by instructors and test developers who use the output of the system. Future work will show the applications with the greatest potential benefit from this approach.

## References


Interagency Language Roundtable. *Interagency Language Roundtable Language Skill Level Descriptions*. Retrieved August 18, 2014, from http://www.govtilr.org/skills/ILRscale4.htm, 2013

Keenan, Thomas, Resnik, Philip, Skaggs, Bradley, & Weinberg, Amy. (2005). *Language Resources for Learning Environments: TTO32 IPR M3*. University of Maryland Center for Advanced Study of Language, College Park, MD

Kincaid, Peter J., Fishburne, Lieutenant Robert P. Jr., Rogers, Richard L., & Chissom, Brad S. (1975). *Derivation of new readability formulas for Navy enlisted personnel*. Research Branch Report 8-75, U.S. Naval Air Station, Memphis, TN.

Petersen, Sarah E., & Ostendorf, Mari. (2009). *A machine learning approach to reading level assessment. Computer Speech and Language*. (Vol. 23, pp. 89-106).

Schwarm, Sarah. E., & Ostendorf, Mari. (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.

Shen, Wade, Williams, Jennifer, Marius, Tamas, & Salesky, Elizabeth. (2013). *A Language-Independent Approach to Automatic Text Difficulty Assessment for Second-Language Learners*. Proceedings of the Workshop on Predicting and Improving Text Readability for Target Reader Populations. Sofia, Bulgaria, pp. 30-38.

Roukos, Salim, Quin, Jerome, & Ward, Todd (2014). *Multi-Lingual Text Leveling*. In Sojka, Petr, Horak, Ales, Kopecek, Ivan, & Pala, Karel (Eds.), To appear in Text, Speech and Dialogue (TSD) 2014. Springer's Lecture Notes in Computer Science (LNCS).