ⴰⵙⵉⵏⴰⴳ-*Asinag,* 9, 2014, p. 105-118

Characterizing the Evolution of Arabic Learners' Texts: A Mostly Lexical Perspective

Violetta Cavalli-Sforza & Mariam El Mezouar

Al Akhawayn University

Nous étudions l'évolution d'une série de textes conçus pour les apprenants de la langue arabe, langue seconde, le long d'un cursus en considérant leur contenu lexical en termes de vocabulaire soi-disant acquis ou en cours d'acquisition par les apprenants auxquels sont destinés ces textes. Nous examinons aussi l'évolution d'autres variables de texte communément utilisés pour mesurer la lisibilité d'un texte. L'objectif est de déterminer les traits des textes qui peuvent être utilisés pour construire un modèle prédictif de la pertinence d'un texte à un apprenant, à un stade d'apprentissage donné, tel que défini principalement par le vocabulaire appris. Nous concluons en examinant si l'approche et les résultats peuvent être appliqués à l'amazighe.

Introduction and Motivation

Reading is one of the four fundamental competences that are targeted when learning a new language, the others being writing, listening, and speaking. The activity of reading serves multiple purposes, of which recognition and understanding of written words in context is a primary one, as it aids in the development of a rich vocabulary and mastery of the nuances of its use. However, a learner cannot read just any text: when creating or choosing a text for language learners, an instructor must consider different goals and constraints. The text must aid in practicing newly learned language concepts—vocabulary and grammar, among others—while at the same time being sufficiently accessible to the learner by containing enough familiar concepts and covering, ideally, an interesting topic. Although the topic of the text can indeed be a motivating or demotivating factor for a learner, except in the case of fully independent and/or rather advanced learners, the choice of topic is usually determined by the instructor in view of the goal of developing specific vocabulary competence. On the other hand, the constraint of "containing enough familiar concepts" cannot be overlooked. In language learning, as in other areas of learning, what is known provides the framework for comprehending and anchoring novelty. A text that contains too many new terms and unfamiliar structures can only be comprehended with great effort, if at all, and will mostly serve to frustrate and demotivate all but the most tenacious of learners.

Novel language concepts are only one aspect of text difficulty; another is the complexity of the text due to the style in which it is written. A text may contain familiar concepts, but they may be expressed in a complex fashion, for example, using difficult words and intricately structured sentences whose relationship to each other is not clearly signaled. Such a text may also require more cognitive effort for successful processing and may lead to frustration and/or failure of comprehension. The problem of readability of a text was addressed initially in the context of learning to read in one's first language, driven by the need to create accessible schoolbooks. Since not all students in a given school grade have the same reading skills, and in some case are weak readers, the complexity of a text can interfere with the learning of subject matter content and work against students success across different subjects. More recently, with extensive information about topics of public interest (e.g., medical conditions and care) becoming available to everybody on the internet, the concern with producing generally and easily readable materials has spread to a much wider range of topics and audiences. Correspondingly, as more text material becomes accessible in electronic format, and as natural language processing (NLP) technology advances, researchers have started turning their attention to the use of computational tools to help evaluate texts for readability and their appropriateness for learners at different levels.

Pioneering efforts in the use of NLP technology for assessing text readability and appropriateness to a learner's level focused on English, followed by other European and oriental languages. Very little work has been done to date on Arabic and, not surprisingly, none on Amazigh, as far as we know. Yet, these two languages are of interest for Morocco today and one characteristic they share is that, in the context of reading, they can be thought of as combining aspects of first and second language learning. Modern Standard Arabic (MSA or فصحة) is no one's mother tongue. Students in Arabic-speaking countries learn MSA at school and, while undoubtedly aided by the dialect in some respects, they are still learning a different and more complex language. The place of MSA as a second language is more evident for individuals who are born in Amazigh-speaking households, and the difficulty of learning to read in MSA is even more acute when considering literacy programs for adults who need to develop character and word-decoding skills for a language that is not entirely the one they usually speak (Maamouri, 2005). Learning to read Amazigh (whether using Tifinagh or the Latin alphabet) poses similar challenges: it is clearly a second language for Moroccans coming from Arabic-speaking households, and the many variants of Amazigh spoken in Morocco alone make the standard form of the Amazigh language used in written educational materials something of a second language even for native speakers.

In view of the dual first-language/second-language position held by a language such as MSA, the work we describe in this paper draws on research performed in both first and second language contexts, and it examines both absolute text readability measures and the appropriateness of a text to a learning stage. Our research focused on MSA, so some of its conclusions are specifically relevant to that language and pertain to linguistic features that are not necessarily present in the Amazigh language. Nonetheless, the general approach remains valid and the results obtained to date for Arabic can inform similar work on Amazigh, thus providing further encouragement to develop NLP technology for this language.

The work reported herein should be viewed as an intriguing but shallow excursion into largely unexplored territory. We are fully aware that there are several aspects of text appropriateness and complexity could not be investigated due to time and resource constraints. We are confident, however, that the work performed and the results obtained to date provide a sound basis for future work.

Readability of a Text vs. Appropriateness of a Text to a Learner

Readability assessment is defined in the literature as the estimation of how 'difficult' a piece of writing is. The assessment is based on the characteristic of the text itself and so is independent of the learner. On the other hand, determining whether a text is appropriate for a specific learner or class of learners requires characterizing the learner(s) and the text in terms of common features and defining criteria that relate the two. In this section we begin by describing a few of the most common and well known readability assessment approaches and measures—which number in the hundreds—and how they have been applied to Arabic. We then examine how a text and a student can generally be characterized in order to establish a way of measuring the appropriateness of the former to the latter. We review some important previous work performed in that area before proceeding to describing our own study.

Measuring Readability

Readability of a text can be defined, as well as measured, in different ways. In 1963, Klare defined it as "a measure of the ease of understanding due the style of writing" (DuBay, 2004); this is a sweeping definition that focuses on the text itself while not precisely referring to any of its specific linguistic or stylistic features. McLaughlin, in 1969, emphasized the relationship between the material and the reader, defining readability as "the degree to which a given class of people finds certain reading matter compelling and comprehensible". A comprehensive definition of readability, which also relates the reader to the text, is given by Dale & Chall (DuBay, 2004) as: "The sum total (including all the interactions) of all those elements within a given piece of printed material that affect the success a group of readers have with it. The success is the extent to which they understand it, read it at an optimal speed, and find it interesting." The complexity and comprehensiveness of these definitions, however, is not reflected in the more quantitative measurement methods developed in the educational literature.

How is readability measured? A widespread method is cloze tests, in which a reader must fill in missing words in the text. Human judgments of readability are also used. However, the most common approaches to measuring readability are formulas that compute a numeric score from some text characteristics, which vary based on the formula. The Fleish-Kincaid Reading Ease Score (FRES), for example, is a linear combination of average word length (measured in syllables per word), and average sentence length (measured in words per sentence) (DuBay, 2004). The Dale-Chall formula uses average sentence length and the ratio of "difficult" words to total words; difficult words are those not present on a list of words expected to be known by a fourth grader in the American school system

(DuBay, 2004). The SMOG Formula uses a somewhat more complex algebraic formulation involving the number of polysyllable words (words with more than 3 syllables) and the number of sentences (Mc Laughlin, 1969). The values obtained from these readability formulas are used to link a text to a grade level, either directly, as in the case of the Flesh-Kincaid Grade Level formula (a variation on the FRES) or the SMOG formula, or by going through a table, as in the case of the Dale-Chall formula. For example, a score of 4.9 and below for Dale-Chall indicates that a text is appropriate for the 4th grade and below, whereas a score of 9 to 9.9 places a text at the college level and a score of 10 and above is for texts suitable for graduate school. These formulas have varying degrees of accuracy but do not typically transfer to languages other than English.

Outside of English, work on readability formulas has been carried out for European languages, such as French, Spanish, Swedish and Danish (Al-Tamimi *et al*., 2013), and for Japanese (Tateisi *et al*., 1988), among others. In addition to the readability formula approach to readability assessment, other approaches are found in the literature for different languages. For example, for Chinese, Pang (2008) describes a method based on Support Vector Machine for regression problems, where key text features are selected and used to predict readability. A study of Hebrew, a Semitic language closely related to Arabic, examined the correlation of 50 features (lexical, semantic, morphological, statistical and syntactic characteristics of the texts) and the difficulty score (1-easiest to 10-hardest) of a set of 70 texts, as rated by language experts (Ben-Simon and Cohen, 2011).

Measuring Readability for Arabic

For Arabic, two formulas were found in the literature: the Dawood Formula (Dawood, 1977), which includes five text features (average word length, average sentence length, word frequency, percentage of nominal clauses, and percentage number of frequency of definite nouns) and the Al Heeti formula (Al-Ajlan *et al.*, 2008), which takes into consideration only the average word length. The selection of these specific features has not been thoroughly justified in the literature, nor do the formulas achieve good results.

We also found two studies in which Saudi researchers used machine learning techniques to learn features for mapping texts to grade levels. In the first of these studies, researchers trained a Support Vector Machine classifier to assign texts, hand-selected from the Saudi school curriculum, to three difficulty levels: easy, medium and hard (Al-Khalifa and Al-Ajlan, 2010). The candidate text features chosen as input to the classifier for each text were: average sentence length, average word length, average number of syllables, word frequencies and perplexity scores for bigram language model. The trained classifier was then tested against unseen texts and expert ratings. The authors concluded that average sentence length was the best single feature in determining Arabic text readability, with a 66.67% accuracy rate; the best combination of features was average sentence length, statistical language model, and term frequency. However, while the prediction accuracy of the model was excellent on easy texts (100%), it dropped to 70% for hard texts and did not achieve any good results on medium texts. The authors attributed the poor results obtained for the latter category to some confounding

factors in the texts themselves. As in previous work (Ajlan *et al.*, 2008), they questioned the validity of the assumption that texts from the Saudi curriculum are correctly distributed among the three difficulty levels, and argued that, to be able to give more accurate predictions from such a system, there is a need for an Arabic corpus correctly labeled with readability levels.

A research effort along similar lines aimed at determining the factors that affect readability of an Arabic text and its mapping to the 10 grades of the Jordanian school curriculum (Al-Tamimi *et al*., 2013). The study was conducted using factor analysis on features that included word length, word frequency, vocabulary load, number of difficult words, average sentence length, sentence complexity, the clarity of the text idea, the use of topology or metaphors, and the grammatical structure complexity. The features were later grouped to remove some redundancy using principal component analysis to determine the most salient features. These were in turn used to create the AARI Base formula, which is then used in another formula to map a text to a grade level. The best performance (with accuracy of over 83%) was obtained when assigning a text to one of three clusters (1st to 3rd grade, 4th to 5th grade, and 6th to 10th grade). The accuracy dropped to under 50% when assigning to individual grades.

A third study took a much simpler approach to assessing the difficulty of texts (Daud *et al*., 2013). It was based on summing the score of words and dividing it by the number of words in the text. The score of individual words is drawn from the frequency of the words in the King Abdulaziz City for Science and Technology Arabic Corpus (KACSTAC) corpus, a general corpus whose texts are derived from magazines, books, newspapers, referred journals, dissertations, government circulation, school curriculums, newswire and the Internet. The authors assume that the more frequent a word is, the easier it is, and so the word score is its reversed ranking in the corpus. Consequently, the lower the overall score of the text, the easier the text is assumed to be. Unfortunately this work appeared to be rather preliminary and did not provide any specific conclusions about the effectiveness of the readability estimation.

Measuring Appropriateness of a Text to a Learner

The readability indices described in the previous section either assign an absolute readability score to a text or relate the text to a grade level in a school curriculum. They presuppose that the grade level is a sufficient, if vague, indicator of an average level of reading skill expected of students in that grade. The actual skills that are expected of a student are characterized, in general, by the learning objectives at each grade level. This characterization is generally taken to be adequate for a gross mapping of texts to curriculum level and for the population of students at that level. It does not really permit any finer level analysis of what the student actually knows, nor what learning is afforded by a specific reading material, and it does not support any adaptation of text selection to the needs of the individual student. In contrast, with advances in artificial intelligence techniques, user-adaptive tools and interfaces are now possible and have become the hallmarks of effective man-machine interaction, whether applied to educational experiences in intelligent tutoring systems or to general information access over the internet.

The development of a more user-adapted interaction between the human user and the machine involves collecting data on users and then selecting from a range of possible information to offer to each user, or group of users, materials or services that seems to best fit their needs at a given point in time. It was with this perspective in mind that the REAP project (http://reap.cs.cmu.edu) developed an approach to text selection addressed at characterizing the student's knowledge and interests, the contents of a text, and the contents of a learning curriculum in such a way as to allow a reading practice environment to select a text to support reader-specific lexical practice (Brown and Eskenazi, 2004). REAP focused on supporting learning English as a second language, but many of its core ideas carry over to learning to read complex texts in one's first language and in literacy training as well. The initial focus on English and vocabulary learning was later extended to cover some grammatical concepts and versions of the system were built for French and European Portuguese.

According to the initial REAP approach, the selection of an appropriate text depends on the vocabulary the student already knows (the *student model*), the vocabulary the student is trying to learn within a curriculum of study (the *curriculum model*), and the vocabulary content of a text (the *text model*). In a nutshell, a text supports reader-specific lexical practice if it contains at least some of the words targeted for learning at a particular curriculum stage, while also containing a sufficient number of known words—and correspondingly a sufficiently small number of unknown words—so as to make the text accessible for the learner. It is this view of what makes a text appropriate or adapted to a learner that underlies the research that we conducted for Arabic and describe below.

Description of the Study

Goals and Approach of the Study

The ultimate goal of our research was to develop a characterization of what makes a text suitable for a learner studying MSA as a second language, so as to be able to predict the suitability of a text to a given *skill stage* and, eventually, a specific learner and dynamically select texts to enrich available learning materials. To our knowledge, there are no freely available corpora of texts for Arabic learners tagged by level of difficulty. So, we used as an implicit measure of increasing difficulty the curriculum presented by part of two volumes of the commonly used Al-Kitaab textbook series (Brustad *et al.*, 2004 and 2007). The small collection of texts contained in this textbook series, each tied to a specific lesson with its vocabulary and grammar concepts, served as our basic data for characterizing the appropriateness of texts to a given skill stage. This characterization or model, once developed, would then serve to inform the selection of further texts to enrich the supply of reading materials available for learners at a specific skill stage. Using the terminology adopted by project REAP, we framed the problem as follows:

- Chapters 1-20 of Al-Kitaab Volume 1 and Chapters 1-5 of Al-Kitaab Volume 2 formed the 25 skill stages of our *curriculum model*.
- Each skill stage had an associated set of vocabulary items explicitly targeted by the chapter as lexical items the student should learn.
- The lexical items presented at each skill stage cumulatively represent the content of a *student model* for the perfect student—the one who does not forget anything—who has reached that skill stage.
- At any one stage in the curriculum, the perfect student is expected to have learned and be able to recognize all words in previous skill stages (the *known* words), is in the process of learning new words for the current skill stage (the *target* words), and is not expected to know any of the words presented in later skill stages (*unknown* words).
- The words contained in the texts and their classification as *known*, *target*, and *unknown* words for a given skill stage are elements of the *text model*, to which we later added other readability variables.

The research examined, qualitatively and quantitatively, trends in the proportions of known, target and unknown words in texts through the 25 skill stages. These variables and some of the standard readability measures were used to train a model to predict which stage an unseen text was appropriate for.

Data Used in the Study

In addition to the Al-Kitaab texts and vocabulary lists that served as the training data for the model, three additional sets of texts were used:

- A set of 23 texts, created or chosen by an Arabic language instructor familiar with the Al-Kitaab book to match specific chapters/skill stages; a few of these were transcription of online videos. These labeled texts were used to test the accuracy of the predictive model.
- A set of 10 texts, hand selected from the Syrian school curriculum (Syrian, 2013) spanning grade levels from primary to high school.
- A set of 10 texts manually collected from the internet from the online newspaper Hespress (http://hespress.com/).

These additional texts, collected using the results of the analyses performed on the Al-Kitaab texts, were intended to be used for prediction from a model. The model was still being built at the time of writing.

Tools and Processing Used in the Study

None of the texts used in the study had any associated information to help identify the words contained in the text. As a result, each text was initially processed by running it through MADA (Habash *et al.*, 2009), which performs several

operations—tokenization, diacritization, morphological disambiguation, part-of-speech tagging, stemming and lemmatization—at the same time. It provides as output a list of ranked analyses for each token in its input. MADA's analyses were aggregated by Buckwalter lemma-ID, a feature that conveys the sense of the word, thereby ignoring analyses that differed only in inflectional features or in the presence of different clitics. MADA is trained on an extensive corpus of texts, of which the Al-Kitaab texts or the other texts we used are not necessarily representative. Evaluation of MADA by its authors on the same type of data on which it was trained shows that MADA's first analysis picks the right part of speech and the right lemma-ID over 96% of the time. Manual examination of MADA's output on 100 randomly selected words in the Al-Kitaab texts showed that MADA was able to pick the right word sense over 94% of the times. Additional analyses we performed using the second MADA result (after aggregation by lemma-ID), shows similar patterns to the first analysis. It was therefore decided that it was safe to base further work on the first MADA analysis.

A MADA analysis was picked for each word in the text and classified as a *known*, *target*, or *unknown* word from the perspective of the perfect student model. A word may have been introduced at different times and/or with different information or expected *competence* in the vocabulary curriculum. For example, the learner may have seen the word in a list of vocabulary targeted by a specific skill stage; if so, this word is labeled as having *high competence*, since the (perfect) learner will supposedly have acquired it before moving on to the next skill stage and will not forget it thereafter. Alternatively, the learner may have been exposed to a word without actually being expected to learn it; for example, the word may have been used in an example with an inline gloss, or in a previous text with no gloss. In this case the learner would not be expected to display high competence. After a number of exploratory analyses, it was decided to use the following procedure to classify words. Let **t** stand for the skill stage at which a text is introduced and **d** stand for the first high-competence appearance of a word in the vocabulary curriculum. At a given skill stage, a word is considered:

Known if **d** is less than **t**

- *Targeted* if **d** is equal to **t**
- *Unknown if* **d** is greater than **t**

Punctuation, number and non-Arabic word tokens were identified and excluded from further analyses. Additional details regarding the extensive data processing and sensitivity analyses performed are described elsewhere (El Mezouar, 2013).

Results of the Study

The major results of the study concern trends in *known*, *targeted*, and *unknown* words in texts and correlation of other common readability variables with skill stages. Progress has also been made towards the construction of a predictive model to be able to assign texts to specific skill stages, but the results obtained, though encouraging, are still preliminary, so this aspect requires further investigation.

Trends in of Known, Targeted, and Unknown Words in Texts

Figure 1 below shows the evolution in proportions of *known*, *targeted*, and *unknown* words in texts across the 25 skill stages. The figure shows the trends for tokens, but very similar patterns were obtained for types (unique word occurrences). While there is a certain amount of oscillation in all word types, there are also clear trends: the percentage of *known* words in texts steadily increases and the percentage of *unknown* words clearly decreases. Moreover, with few exceptions, the number of unknown word types is quite small, never going above 46 and usually below 20. These results confirm general expectations.

Past the first 3 skill stages, where a significant percentage of new vocabulary is used in short texts, *targeted* words hold steady within a range, 2-19% for tokens and 2-20% for types, with an average of 8% of targeted words for both tokens and types. It is an open question whether these values reflect an intentional choice by the textbook authors, who created or selected the texts or whether they represent an upper bound on how many new words can reasonably be targeted in a single text.

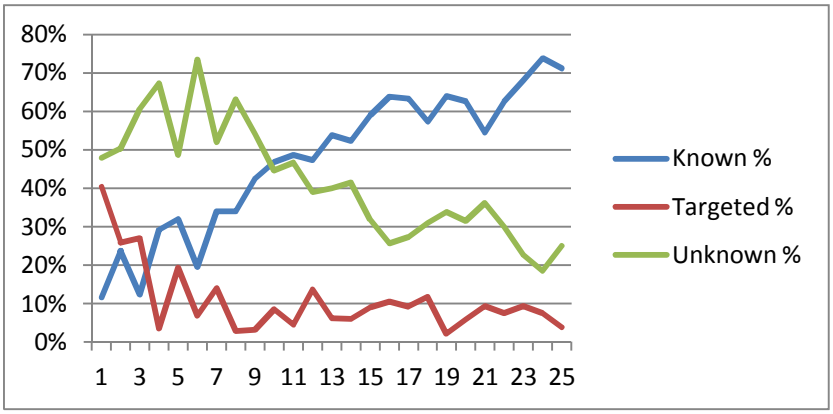

Figure 1 : Evolution of known, targeted and unknown variables for tokens

In addition, we investigated the vocabulary load across different skills stages and remarked there was a definite fluctuation, from low to high rates of new vocabulary introduction, with a periodicity varying between 2 and 4 skill stages. This indicates that there are points in the curriculum in which there is more emphasis on new vocabulary and others where the vocabulary is used and reinforced through practice.

Correlation of Common Readability Variables with Skill Stages

We considered some of the other measurements of text complexity that appear in common readability indices for Arabic and other languages to determine if any could be valuable indicators of the appropriateness of a text to a given skill stage. Table 1 summarizes some of the results obtained.

Particularly interesting is the negative result obtained for average length of words, a variable that appears, in some guise or other, in readability formulas for other languages, where it may measure number of characters or syllables. We note that word formation processes for Arabic, with 'word' defined as a series of characters bounded by spaces or punctuation symbols, differ significantly from those of most of the languages for which readability measures we reviewed were defined. Specifically, the Arabic word becomes longer on account of three processes:

Text feature	**How computed**	**Good determinant?**
Familiarity with vocabulary	Percentages of targeted, known and unknown words in a text	Yes
Open-class words	Percentage of open class words in a text	Yes
Closed-class words	Percentage of closed class words in a text	No
Lexical diversity	Ratio of unique words over total number of words in a text	No
Length of texts	Number of tokens in a text	Yes
Average length of words	Average number of characters per word in a text	No
Average complexity of words	Average number of clitics per word in a text	Yes
Average length of sentences	Average number of words per sentence in a text	Yes

Table 1: Summary of features determining appropriateness of text to level

1. **Derivational Processes**: intersection of the root with longer patterns containing more fixed letters. The different patterns are associated with different meanings and there are a fixed and small number of patterns and character additions, so the reader quickly develops skill in identifying these components of the word. Examples include the triliteral verb measures V and X, which add a 't' (ت) and an 'ist' (است) prefix, respectively. However the amount of word lengthening due to this process is rather limited.
2. **Inflectional Processes**: the addition of inflectional prefixes and suffixes, such as the markers for person and number in verbs and for gender, dual and plural in nouns and adjectives. Again, the amount of word lengthening due to this process is rather limited in both amount and variation.
3. **Cliticization**: The prefixation or suffixation of morphemes, including conjunctions, prepositions, pronouns and other particles, depending on the position. There are up to 4 proclitics (prefixed) and 1 enclitic (suffixed).

In light of the above word-formation processes, it is not surprising that, for Arabic, counting clitics gives more meaningful results than counting letters.

While we are encouraged by the positive results (features that are good determinants), we are also aware of the fact that some of the negative results may be due to the limits of the processing we performed, as described below.

Limitations of the Study

The initial motivation for this work came from a project targeted at developing Arabic reading enhancement tools (Maamouri *et al.*, 2012). That project was itself a response to some of the special difficulties involved in reading in MSA, given that certain aspects of its morphology and its writing system conspire to complicate significantly the recognition of a word in context. Firstly, as noted earlier, Arabic morphology allows the adjoining of several prefixes and suffixes to the basic word stem, so that identifying the stem involves stripping off affixed material, a process that suffers from some ambiguity. Secondly, the stem itself may be difficult to identify because of internal or boundary changes, such as those occurring in broken plurals or in words whose roots contain weak consonants. Thirdly, traditional Arabic dictionaries are organized by root, so looking up a word requires identifying the root letters and the pattern that give rise to the stem, a process that is also complicated by the presence of weak consonant radicals and assimilation processes. Finally, the absence of most diacritic signs in all texts, other than religious texts or texts used in early school years, further enhances the ambiguity of words and contributes to the difficulty in their identification. All these factors contribute to making word recognition in context quite difficult for the learner of MSA, since it requires extensive application of different sources of linguistic knowledge—lexical, morphological, syntactic, and semantic—in addition to general common sense knowledge and topic knowledge.

While the data we worked with contained some morphosyntactic information we could have used, our focus for this study was limited to words and, more particularly, to word senses (as captured by lemma-ID), treating words that share the same sense as equal, independently of the form they took on in context. In so doing, we ignored morphological processes giving rise to different inflectional and derivational variants. Similarly, we did not consider those closed-class words that appeared as clitics attached to an open-class word. We also labeled as unknown words whose meanings were not explicitly presented in the vocabulary but should or could have been guessed by the learner using acquired knowledge of morphological processes covered in the grammar. Investigating the effect of these and other omissions is left for future work.

Conclusion and Implications for Amazigh

We have described an exploratory study that investigated certain aspects of texts assumed to be relevant in determining the appropriateness of a text to a learner at a certain skill stage. We analyzed a text from two perspectives: 1) its fit to the lexical knowledge the learner might (or might not) have and the vocabulary targeted for

learning at a given stage in a curriculum; and 2) specific characteristics of the text, independent of the learner or of the instructional curriculum. The latter characteristics were chosen among the ones used in common readability measures for different languages. Most of the analysis focused on the lexical content of texts, and specifically the words that a learner was supposed to have acquired before reading a text and those contained in the text to aid the learner in practicing new vocabulary. We were able to find definite trends in these measures. We were also able to verify that some of the variables used to measure text complexity in well-known readability indices were informative as determinants of appropriateness to a learner's skill level, while others did not appear to be. However, it was also remarked that some of the negative results for some promising variables—for example, lexical diversity and proportion of closed-class words—may be due to the limited processing we performed and the morphosyntactic information it ignored. A fuller analysis that takes these characteristics into consideration could yield different answers. Indeed, other simple measures of morphological and syntactic complexity, such as average complexity of words—measured in number of clitics—and sentences—measured in average words per sentence—suggest that this information might be more useful than immediately apparent.

The ultimate goal of the analysis reported herein was to build a model that would allow us to predict whether a given text could be appropriate for a given skill stage, that is, would be accessible to a perfect learner who had learned all of the vocabulary presented up to that stage and would be adequate for practicing the use of lexical items introduced at that stage. The predictive model is still under development but preliminary results and feedback from instructors provide grounds for optimism. Further refinements and investigation are required to make the model truly useful and will be the focus of future work.

In general, it is a worthwhile goal to build such a model and use it as part of a language learning tool to provide criteria for (semi-)automatically selecting and/or modifying texts for learners, in order to dynamically enrich a database of learning materials. The general approach followed for Arabic in this study can be applied to other languages as well. Nonetheless, just as it has been shown that standard readability indices do not transfer directly between languages, we do not expect our analysis and the specific results we obtained to be applicable to all languages. With respect to Amazigh, we can expect that familiarity with vocabulary will be relevant to choosing learner-appropriate texts, as will open-class words, length of texts and average length of sentences. On the other hand, we have to withhold judgment on features such as average length of words and average complexity of words, since Amazigh morphology has some similarities to Semitic morphology but also many differences. Specifically, it is significantly less "agglutinating" than Semitic language morphology, even though Semitic languages are not themselves considered agglutinating languages (as opposed to, for example, Turkish). Concerning the significance of other features, such as the proportion of closed-class words or lexical diversity, more detailed analyses need to be performed even for Arabic, and it may well be that Amazigh will turn out to behave similarly to other languages for which these features do play a role in readability of texts. One aspect of morphology that we have not yet investigated may turn out to be important for text difficulty in both MSA and Amazigh: the change in context of

certain radical letters, e.g. the 'w/u' and 'y/i' sounds as well as other assimilation processes that are known to occur in both languages in the presence of adjacent consonants.

As a concluding remark, we point out that a study of the complexity and evolution of learners' texts along a curriculum, such as described above for MSA, is really enabled by the presence of digital text corpora (annotated or not), computational lexicons, part-of-speech taggers and morphological analyzers, among others. Therefore, the advancement of language instruction for both first and second language learners of Amazigh provides yet more motivation for investing in the development of such resources for this language.

Acknowledgements

This research was inspired and made possible by a previous project, housed at the Linguistic Data Consortium (LDC), and supported by the U.S. Department of Education under International Research Study (IRS) Grant No. P017A050040-07-05. We also thankfully acknowledge the support and generosity of the Linguistic Data Consortium for allowing us to use some of their linguistic resources. For further information regarding that project, contact Dr. Mohamed Maamouri (PI) at maamouri@ldc.upenn.edu.

References

Al-Ajlan, A. A., Al-Khalifa, H. S., and Al-Salman, A. S. (2008), « Towards the Development of an Automatic Readability Measurements for Arabic Language », *in Proceedings of the Third International Conference on Digital Information Management*, November 13-16, p. 506-511.

Al-Khalifa, H. S., and Al-Ajlan, A. A. (2010), « Automatic Readability Measurements of the Arabic Text: An Exploratory Study », *The Arabian Journal for Science and Engineering*, Volume 35, Number 2C, p. 103-124.

Al-Tamimi, A-K., Jaradat, M., Aljarrah N., and Ghanem, S. (2013), « AARI: Automatic Arabic Readability Index », *The International Arab Journal of Information Technology*, [Accepted March 12, 2013].

http://www.ccis2k.org/iajit/PDF/vol.11,no.4/5200.pdf, August 2013.

Ben-Simon, A. and Cohen, Y. (2011), *The Hebrew Language Project: Automated Essay Scoring & Readability Analysis*, International Atomic Energy Agency.

http://www.iaea.info/documents/paper_4e1237ae.pdf, August 2013.

Brown, J., and Eskenazi, M. (2004), « Retrieval of authentic documents for reader-specific lexical practice », *in Proceedings of InSTIL/ICALL Symposium*, June 17-19, Venice.

Brustad, K., Al-Batal, M., and Al-Tonsi A. (2004), *Al-Kitaab fii Ta^{c}allum al-^{c}Arabiyya, A Textbook for Beginning Arabic: Part One*, Second Edition, Washington D.C., Georgetown University Press.

Brustad, K., Al-Batal, M., and Al-Tonsi A. (2007), *Al-Kitaab fii Ta^{c}allum al-^{c}Arabiyya, A Textbook for Beginning Arabic: Part Two*, Second Edition, Washington D.C., Georgetown University Press.

Daud, N. M., Hassan H., and Abdul Aziz, N. (2013), « A Corpus-Based Readability Formula for Estimate of Arabic Texts Reading Difficulty », *World Applied Sciences Journal*, Volume 21, p. 168-173.

DuBay, W. H. (2004). *The Principles of Readability*. Institute of Education Sciences, http://eric.ed.gov/?id=ED490073, August 2013.

Dawood, B. A-K. (1977). *The Relationship between Readability and Selected Language Variables*. Thesis submitted to the College of Education Board in the Universit of Baghdad in partial fulfillment of the requirements for the degree of Master of Arts in Education and Psychology.

http://dspace.ju.edu.jo/xmlui/bitstream/handle/123456789/12718/JUA0305740.pdf, August 2013.

El Mezouar, M. (2013), *Appropriateness of a Text for Learners of Arabic as a Foreign Language: A Word-Based Perspective*, Master Thesis Report submitted in partial fulfillment of the requirements for a Master of Science in Software Engineering at the School of Science and Engineering of Al Akhawayn University in Ifrane.

Habash, N., Rambow, O., and Roth, R. (2009), « MADA+TOKAN: A Toolkit for Arabic Tokenization, Diacritization, Morphological Disambiguation, POS Tagging, Stemming and Lemmatization », *in Proceedings of the 2nd International Conference on Arabic Language Resources and Tool*s (MEDAR), April 22-23, Cairo, 2009.

Maamouri, M. (2005), « Arabic Literacy», *in Encyclopedia of Arabic Language and Linguistics*, Lemma 11,16, Volume 2, Brill.

Maamouri, M., Zaghouani, W., Cavalli-Sforza, V., Graff, D., and Ciul, M. (2012), « Developing ARET: An NLP-based Educational Tool Set for Arabic Reading Enhancement », *in Proceedings of The 7th Workshop on Innovative Use of NLP for Building Educational Application*s (NAACL-HLT-2012), June 7, Montreal, p. 127-135.

Mc Laughlin, H. G. (1969), « SMOG Grading - a New Readability Formula », *Journal of Reading*, Volume 12, Number 8, p. 639-646.

Pang, L. T. (2006), *Chinese Readability Analysis and its Applications on the Internet*, Thesis submitted in partial fullfilment of the requirements for the degree of Master of Philosophy in Computer Science and Engineering, The Chinese University of Hong Kong.

Syrian (2013), « المناهج الجديدة », http://me.syrianeducation.org.sy/ebook/classes.html, March 2013.

Tateisi, Y., Ono, Y., and Yamada, H. (1988), « A Computer Readability Formula of Japanese Texts for Machine Scoring », *in Proceedings of COLING 1988*, Volume 2, August 22-27, Budapest, p. 649-654.