Extensions to a Histogram-Based Student Modeling Approach to Facilitate Reading in Morphologically Complex Languages

Violetta CAVALLI-SFORZA
Language Technologies Institute
Carnegie Mellon University
violetta@cs.cmu.edu

Mohamed MAAMOURI
Linguistic Data Consortium
University of Pennsylvania
maamouri@ldc.upenn.edu

Abstract. We propose an approach to student modeling in the context of a project aimed at aiding readers negotiate authentic texts in languages where reading is particularly difficult due to the morphological complexity of the language, among other factors. We focus on Modern Standard Arabic as an example of such a language. Our approach extends existing tools for modeling reading skills, text difficulty, and curricula developed for English. We explore the extensions necessary for supporting Arabic morphology.

Introduction and Objectives

In this paper we describe our intended approach to student modeling for language tutoring in the context of a project titled "Teaching and Learning Linguistically Complex Languages", recently funded by the United States Department of Education under the Title VI International Research and Studies Program. The project aims to support foreign language learning and to enhance cross-cultural understanding by producing substantive textual and lexical learning materials and computer-based instructional tools that aid learners in reading authentic materials in languages that present special difficulties for reading. The specific goals of the project are:

(1) Providing readers with tools to negotiate the complex morphology of target languages;
(2) Enabling learners to read authentic texts containing unfamiliar and difficult words;
(3) Enabling teachers to prepare texts for classroom use and to test students' reading ability; and
(4) Creating easy Internet access to all tools and materials for teachers and learners.

While the tools themselves will be designed to address multiple languages, they will be implemented specifically to support Modern Standard Arabic (MSA), a less-commonly taught but critical language of high priority in Middle Eastern and North African studies. MSA's writing system and morphosyntactic structure present special challenges for the reader, particularly with respect to word identification and lookup in a dictionary. The planned tools address dictionary lookup, text preparation, and assessment of word recognition. To substantiate the claim that the tools do indeed generalize beyond MSA, they will be evaluated with a second less-commonly taught language, Nahuatl, spoken in southern United States and northern Mexico, which presents comparable – though different – word identification and lookup challenges. For MSA, where we have access to substantial textual resources, we will also use the REAP technology, developed at Carnegie Mellon University's Language Technology Institute, to intelligently select texts to be presented to readers based on models of curriculum, text difficulty, student reading skills, and possibly topic interest. The REAP tools were originally developed to improve reading skills through individualized reading practice in English as a first or foreign language and will need to be extended to account for the special challenges presented by reading Arabic texts.

Work on the project will only begin in July of 2005, therefore this paper and our participation in the workshop has two primary objectives: 1) to describe the problem we are attempting to solve and the tools and approaches we are planning to use; and 2) to elicit feedback and learn from other workshop participants with respect to the student modeling component of our reading facilitation tools. Our proposed approach to student modeling, which is heavily based on REAP's histogram-based approach, does not attempt to address all aspects of language learning. Rather, it focuses on modeling specific aspects of reading skill in languages where even the basic process of word recognition presents special challenges due to the writing system and/or the morphosyntax of the language.

1. Reading Arabic Texts: Challenges and Tools

Modern Standard Arabic (MSA) is the primary, if not the only, formal written language used throughout the Arab world and is classified at the highest level of difficulty (level 4) in the United States Foreign Service Institute chart, requiring longer times for mastery than many other languages. Reading in Arabic presents special challenges due to its script. Learners of MSA – the main focus of Arabic teaching in the U.S. and elsewhere, and the only form of written Arabic – face difficulties in word recognition, word disambiguation, and the acquisition of decoding skills, which are important components of reading skill [1] [2]. Authentic Arabic texts lack short vowels and other diacritics that distinguish words and mark grammatical functions. Moreover, Arabic has a rich inflectional and derivational morphology that adds prefixes and suffixes and alters the stem of words according to syntactic context, and utilizes a number of particles (conjunctions, prepositions and pronouns) that attach to words as prefixes and suffixes.

The aforementioned linguistic complexities result in significant reading difficulties. In order to understand the precise meaning of a text, learners who are trying to read materials must insert short vowels and other diacritics themselves on the basis of limited vocabulary knowledge and on the basis of grammatical rules they have not yet completely internalized. To accomplish this, they must be able to recognize letter and word boundaries, decode unvocalized words, and identify and comprehend these words. For example, the word f^{c}lm/ فعلم is composed of the particle fa/ف and one of several possible words written as ^{c}lm/علم (such as: ^{c}ilm "science or knowledge", ^{c}alam "flag", ^{c}allam " (he) taught", ^{c}ulima "it was learned"), and it may play a different role in the sentence depending on unwritten vowels and other diacritic signs. Learners must bring knowledge of vocabulary, root-and-pattern morphology with complex derivational and inflectional rules, syntax, and contextual interpretation to produce correct and meaningful vocalization, to reach final word recognition, and even to look up a word in an MSA dictionary. It is worth noting that MSA presents reading difficulties even for schoolchildren in Arabic speaking countries. Their native language is a spoken dialect of Arabic, whose pronunciation, vocabulary and syntax can differ widely from MSA. Often MSA is their first written language and their first second language (in some areas of the Maghreb region, French has played this role at times). Arabic schoolbooks begin with almost full diacritics and gradually decrease their use through the school years until they are entirely omitted by the end of middle school.

Arabic instruction is challenging for teachers and institutions as well as for learners. Though lately there has been an increasing demand for Arabic instruction in the U.S., and more educational institutions are beginning to offer introductory courses in Modern Standard Arabic (MSA) and some Arabic dialects, Arabic is still not a widely taught language. At higher levels of instruction, there is a shortage of pedagogically sound instructional materials and an insufficient number of teachers who have both the linguistic and technical skills and time to develop such resources, yet exposure to accessible, motivating and authentic materials is key in language learning. Technology is therefore increasingly being used to supplement the model of the teacher-fronted classroom and to foster learner autonomy by adapting instruction to the needs of individual students who may have specific career or academic objectives that require more rapid attainment of advanced language proficiency for better cross-cultural understanding.

To address the above challenges, our project will develop the following tools together:

(1) **Dictionary Lookup Tool:** enables language learners to look up the citation form of an arbitrarily inflected word in a morphologically complex language;

(2) **Reading Facilitation Tool:** enables language learners who encounter an unfamiliar word in electronic text to easily obtain a morphological analysis of that word, together with the dictionary entry for the citation form of that word;

(3) **Word Recognition Assessment Tool**: aids in the assessment of learners' reading ability, specifically the ability to choose the correct morphological analysis (and in languages where this is relevant, the diacritics) for each word, and its corresponding English gloss;

(4) **Text Preprocessing Tool:** designed to help teachers produce texts for use in the Word Recognition Assessment Tool.

The four tools will make use of the Buckwalter Morphological Analyzer, which has been partially developed at and is currently distributed by the Linguistic Data Consortium (LDC) at the University of Pennsylvania (www.ldc.upenn.edu). The project also leverages LDC's extensive and expanding Arabic language resources and in particular the Penn Arabic Treebank to provide a large database of texts for learners and teachers to chose from. The mission of the LDC is to continually collect and make available to the scientific community large quantities of linguistic resources, both text and speech, for Arabic and other languages. In addition to large quantities of raw Arabic text (LDC currently has more than 600 million-words of newswire text and adds 80 million words annually to its collection), it has already published three segments of the Arabic Treebank, with a fourth one close to completion. The treebank contains morphologically and syntactically annotated MSA text including newswire from the Agence France Presse, and the middle eastern newspapers *Al-Hayat* (distributed by Ummah Arabic News Text), *An-Nahar*, and most recently the Tunisian daily *Assabah*.[1] Several more segments of the Arabic Treebank are planned.

2. Supporting Reading Progress with REAP

In addition to building the aforementioned four tools to support reading practice, creation of prepared texts, and word recognition assessment, we plan to interface them with technology

[1] By end of Spring 2005, the Arabic Treebank will contain a total of 791,681 tokens representing about 1 million words after cliticization. The annotated corpora include complete vocalization including case endings, lemma IDs, more specific part-of-speech tags for verbs and particles, and an English gloss for each word.

developed by project REAP (http://orleans.lti.cs.cmu.edu/Reap/) to intelligently select texts for readers from an existing pool of materials [3] [4] [5]. REAP is funded by the U.S. Department of Education and includes researchers from Carnegie Mellon University's Language Technology Institute and the University of Pittsburgh's Learning Research and Development Center. The project aims to find appropriate authentic documents for students learning to read. It shares with our project the concern that too often students are given prepared texts, which has two disadvantages: first, the student is not exposed to examples of real language, that is, the language used in everyday written communication; second, the students all get the same texts to read, regardless of individual reading skills and interests. The REAP project, which was motivated by the desire to improve reading skills through individualized reading practice in the context of an English and ESL classroom/curriculum, is based on L1 reading research, but can be used for L2 reading as well. REAP has developed tools to a) retrieve texts from the Internet or from pre-existing collections that match different curriculum levels, b) model students' reading ability, and c) select texts that are suited to students' reading ability but also move them towards a higher level of reading skill (as defined by the curriculum) and/or pertain to topics of interest to the student or the teacher's lesson focus.

2.1 The REAP Approach to Student, Text, and Curriculum Modeling

There are four types of models in REAP: a curriculum model, two kinds of student models, and text models. REAP defines a reading curriculum with degrees of text difficulty in terms of vocabulary that a student should know at different curriculum levels. The student's knowledge is modeled as two histograms of words: 1) the **passive model**, which consists of all the words the student has read using the system, along with word frequencies – this can be considered exposure to words; and 2) the **active model**, which includes only the words for which the student has somehow demonstrated knowledge. Finally, texts are modeled by a histogram of word frequencies.

In order to present the reader with appropriate texts, a search engine is first used to look for texts that match that curriculum level/reading difficulty and may include other criteria, such as topic-specific vocabulary. For English REAP, the search is performed offline over the web, but it can also performed on a limited collection of texts in real time. To match documents to a student's level, the system then looks at words in the student's active and passive model and the words in the retrieved documents, selecting those documents that contain some subset of known words and some percentage of new words (the stretch). Stretch size can be experimentally manipulated. Once a set of documents appropriate to the student's reading level has been selected, they can also be ranked according to other criteria, e.g. words the student doesn't know but should in order to achieve curriculum level, or frequency of occurrence of these words, or topic of interest for a particular lesson.

2.2 Extending REAP for Arabic

The REAP project tools were developed primarily with English in mind. REAP currently uses unknown vocabulary, excluding named entities, as the sole criterion for modeling curriculum, student knowledge and text difficulty, although some extensions may be undertaken for other linguistic phenomena, and especially English constructions. The bare word models are extended with part-of-speech information. Words with multiple POS are considered different words and, in fact, word cohorts – e.g., ‘read’ ‘reading’ ‘reader’ – raise issues in choosing documents for the student. This is currently a topic of research, to which experience with MSA’s complex derivational morphology can contribute. Knowledge of vocabulary is certainly very important for Arabic learners as well, but must be modulated by other considerations. Morphologically, English is a (relatively) impoverished language, so a number of extensions will be needed in order to capture those aspects of Arabic writing and morphosyntax that make it difficult to decode and identify words and understand the role they play in a sentence. We envision the following major differences and extensions in the treatment of curriculum, text, and student models when applying REAP tools to Arabic.

Treatment of Named Entities: In English, names seldom affect comprehension. In Arabic, however, where there is no capitalization to distinguish proper nouns from regular words, identifying named entities is an important part of word recognition and text comprehension. Many adjective and noun forms are used as names, and their identification as proper nouns depends on knowledge of morphology and syntactic structure. A further problem is posed by the transliteration of foreign names into Arabic script: sometimes the resulting words are easily identified as foreign because they do not fall into the inflectional/derivational patterns of Arabic, but sometimes they do. To what extent it is desirable or feasible to model this problem remains to be determined and is likely to be of secondary priority: the best strategy could well be, at least initially, to make evident in the texts their special nature of named entities (they are specially tagged in the Arabic Treebank), allowing readers to focus on more general and pervasive morphosyntactic phenomena.

Modeling of Morphological Knowledge: Curriculum, texts, and student models, and the tools that operate on them, will need to be augmented with knowledge of inflectional morphological patterns. At this stage of our thinking, such patterns are best represented as collections of morphological features, including part-of-speech, and their surface realization for different categories of words, notably derivational patterns and words containing weak consonants. Included in morphological knowledge categories will be those closed parts of speech that attach themselves to words (e.g. direct object pronouns, conjunctions and prepositions), their effects on the surface realization of words (e.g. the preposition ‘ل’ causes an initial ‘ا’ to be elided), and constraints governing their attachment,. Modeling of derivational morphology skills (as exemplified by the patterns ‘teach’, ‘teacher’ in English, ^{c}allam and mu^{c}allim in Arabic) will need to be left for later, since the Arabic electronic lexicon and morphological analyzer underlying the tools are stem-based and do not attempt to recover derivations from Arabic roots.[2] There is wisdom in using stem-based lexicons: while derivational patterns are quite regular, their accompanying derived meanings are often not.

Modeling Syntactic Context: To the extent made possible by the Arabic Treebank syntactic representation, we will model the syntactic context that affects morphological realization of words. While we do not expect to be able to cover the entire grammar, we will be able to model certain (local) phenomena, for example the omission of the definite article in all but the last term of a construct state (‘iDafa’), or the rule that a verb preceding its subject does not need to agree with it in number (and even not in gender).

Updating the Active Student Model: While the passive student model can be updated by considering which words and morphosyntactic structures are present in texts the students have been exposed to, the active student model must be updated based on the knowledge demonstrated by the student. In REAP’s use with English, knowledge is demonstrated by answering a question about a word; for Arabic, we will need to obtain this information from the Word Recognition Assessment Tool and/or the Reading Facilitation Tool.

[2] Ongoing work may however make this possible at a later date [6] [7].

The use of REAP tools with a morphologically complex language such as Modern Standard Arabic gives rise to an exciting synergy between projects. On one hand, REAP tools will aid the proposed LDC tools to select texts for learners according to pedagogically sound criteria; project team members will interact with language teachers at the University of Pennsylvania and the University of Pittsburgh to develop a curriculum defining levels of text difficulty. On the other hand, the addition of morphosyntactic analysis in modeling the curriculum, text difficulty, and learner ability, provides an opportunity to extend REAP tools in ways not afforded by their application to the English language alone.

3. Background and Qualifications of Authors

Neither of the authors is an expert in the area of student modeling per se, however they both bring relevant and complementary skills to the task. **Violetta Cavalli-Sforza** is a Visiting Researcher at Carnegie Mellon University's Language Technology Institute (CMU-LTI). As a doctoral student in Intelligent Systems at the University of Pittsburgh, and as a staff member and researcher at CMU-LTI, she worked on different aspects of tutoring systems and natural language processing. Her most recent research has focused on machine translation and Arabic morphology generation [6] [7] [8], some of which is being performed in Morocco through National Science Foundation and Fulbright fellowships. She is fluent in four languages, has studied a few more, is a permanent student of Arabic and is well acquainted with the difficulties in learning to read MSA. **Mohamed Maamouri** is a Senior Research Administrator and head of the Arabic Treebank project at LDC. Maamouri is a recognized Arabic language specialist, with significant experience in Arabic reading research, literacy research, and foreign language teaching and learning pedagogy [9] [10]. For over fifteen years he was the director of the Bourguiba Institute of Modern Languages in Tunisia where he started the well-known MSA summer intensive courses. Subsequently he worked as a senior researcher and the associate director of the International Literacy Institute, in the Graduate School of Education at the University of Pennsylvania. For the past three years, Maamouri has been leading the Arabic projects at LDC, where he has overseen and managed the preparation of extensive annotated corpora in Arabic.

References

[1] Perfetti, C. A. (1986). *Reading Ability*. Oxford University Press. New York.

[2] Perfetti, C.A & Hart, L. (2001). The Lexical Quality Hypothesis. In Verhoeven, L., Elbro, C. & P. Reitsma (Eds.), *Precursors of functional literacy*. John Benjamins. Amsterdam/Philadelphia.

[3] Brown, J. & Eskenazi, M. (2004). Retrieval of Authentic Documents for Reader-Specific Lexical Practice. In *Proceedings of InSTIL/ICALL Symposium*. Venice, Italy.

[4] Collins-Thompson, K. & Callan, J. (2004). Information Retrieval for Language Tutoring: An Overview of the REAP Project (poster description). In *Proceedings of the Twenty Seventh Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Sheffiel*d, UK.

[5] Collins-Thompson, K. & Callan, J. (2004). A Language Modeling Approach to Predicting Reading Difficulty. In *Proceedings of the HLT/NAACL 2004 Conference*. Boston, USA.

[6] Cavalli-Sforza, V., Soudi, A., & Mitamura, T. (2000). Arabic Morphology Generation Using a Concatenative Strategy.” In *Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics* (NAACL 2000), 86-93. Seattle, USA.

[7] Cavalli-Sforza, V. & Soudi A. (2003). Enhancements to a Morphological Generator to Capture Arabic Morphology. In *Proceedings of the Eighth International Symposium on Social Communication*. Center of Applied Linguistics, Santiago de Cuba, Cuba.

[8] Soudi, A., Cavalli-Sforza, V., & Jamari, A. (2002). A Prototype English-to-Arabic Interlingua-based MT System, In *Proceedings of the Workshop on Arabic Language Resources and Evaluation – Status and Prospects*. Third International Conference on Language Resources and Evaluation (LREC 2002). Las Palmas de Gran Canaria, Spain.

[9] Maamouri, M. (1998). *Language Education and Human Development: Arabic Diglossia and its Impact on the Quality of Education in the Arab Region*, Preliminary Copy. Discussion paper prepared for The World Bank, The Mediterranean Development Forum, Marrakech, 3-6 September 1998.

[10] Maamouri, M. (forthcoming). Arabic Literacy. In *Encyclopedia of Arabic Language and Linguistics*, Vol. 2. Brill Academic Publishers. Leiden, The Netherlands.