ISCA Archive
http://www.isca-speech.org/archive

Speech and Language Technology in Education (SLaTE2009)
Wroxall Abbey Estate, Warwickshire, UK
September 3-5, 2009

Porting REAP to European Portuguese

Luís Marujo[1], *José Lopes*[1], *Nuno Mamede*[1], *Isabel Trancoso*[1],
Juan Pino[2], *Maxine Eskenazi*[2], *Jorge Baptista*[3], *Céu Viana*[4]

[1] INESC-ID Lisboa / IST, Portugal, [2] LTI / CMU, USA
[3] Univ. Algarve Portugal, [4] CLUL, Portugal
Luis.Marujo@inesc-id.pt

Abstract

This paper describes the early stages of porting REAP, a tutoring system for vocabulary learning, to European Portuguese. Students learn from authentic materials, on topics of their preference. A large number of linguistic resources and filtering tools have already been integrated into the ported version. We modified the current system to also target oral comprehension.

1. Introduction

REAP stands for "Reader-Specific Lexical Practice for Improved Reading Comprehension". It is a tutoring system developed at the Language Technologies Institute at Carnegie Mellon University that supports language teaching to either native or non native speakers. It focuses on vocabulary learning by presenting readings with target vocabulary words in context to students [1].

In the REAP platform, students can learn from authentic materials selected from an open corpus such as the Web, on topics for which they previously marked their preference. The passages are also selected to satisfy very specific lexical constraints, and are suited to each student's degree of acquisition and fluency for each word in a constantly-expanding lexicon.

The development of a Portuguese version of REAP (REAP.PT) is one of the goals of the CMU-Portugal dual PhD program in the area of language technologies, which brings together an interdisciplinary team of engineers from the Spoken Language Systems Lab of INESC-ID Lisboa, and linguists from the Universities of Algarve and Lisbon.

Porting REAP to a new language involves several tasks, the most obvious one being the adaptation of the interface to Portuguese. But it also involves integrating new linguistic resources (such as dictionaries for looking up the meaning of unfamiliar words, web corpora, target word lists), new tools (such as topic classifiers), and making the necessary adaptations for this typologically different language.

There are several other tasks, however, where the project goals are not limited to porting, and involve research challenges. Readability is one such task, where the aim is to attribute a reading difficulty value to a document. Cloze question generation is another task aiming at automatically generating fill-in-the-blank questions with multiple choices, which are used as practice exercises for focus words.

In addition to these porting and research-oriented tasks, REAP.PT is also concerned with the integration of oral comprehension features in the tutoring system. Learning a new word does not only mean learning how to write it but also how to pronounce it. This may be especially important for a language such as European Portuguese, characterized by strong vowel reduction (ranging from quality change to shortening and deletion), where familiarization with the spoken/written language in multimedia documents may turn out very helpful for non-native speakers.

This paper addresses the early stages of porting to Portuguese, emphasizing the large number of linguistic resources that the ported version already integrates (Section 4), the successive document filtering stages (Section 5), and the modifications introduced to target oral comprehension (Section 6).

2. Related work

Several authors agree on the division of the progress of CALL (Computer Aided Language Learning) into three phases known as "Behaviorist", "Communicative" and "Integrative". The REAP system falls into this latter phase which is centered in Multimedia and Internet, resulting from the shift to globalization, where the teachers turn into facilitators instead of being the source of knowledge and the students should interpret and organize the information given in an active way.

A number of recent projects have taken similar approaches to providing language learners with authentic texts. WERTi [2] is an intelligent automatic workbook that uses texts from the Web to increase knowledge of English grammatical forms and functions. READ-X [3][4] is a tool for finding texts at specified reading levels that also performs a classification per area of interest. SourceFinder [5] is an authoring tool for finding suitable texts for standardized test items on verbal reasoning and reading comprehension.

REAP takes a slightly different approach. Rather than teachers choosing texts, the REAP Tutor itself selects individualized practice readings from a digital library. The readings contain target vocabulary words that a given student needs to learn based on a student model.

3. System architecture

Students interact with REAP via a web interface, supported by any web browser available. At the first login, the tutor gives a pre-test, in which the interface shows the target word list, and asks the students to choose the ones they know, in order to assign one of the 12 school levels. After the pre-test, the student menu displays several options: group readings, individual readings, oral comprehension, and topic interests. The first two are very similar, the only difference being that in the first case the text selected for reading is chosen by the teacher and is common to all the students in the class. The topic interest menu displays a list of topics and asks the student to classify them by checking one of five boxes from "not interested" to "very interested", storing the information in the database.

During a reading session, the focus words are highlighted in the texts and the student can search for the meaning of the words by clicking on them or by using the search field of the system. This is important because we want to track every action by the student, namely the access to the dictionary, in order to keep updating the progress of the student. The reading session is followed by a series of cloze, or fill-in-the-blank, questions about the words that were highlighted.

The interface also has a teacher menu that allows the teacher to rate the quality of a document, estimate the readability level, select documents for group reading, discard documents, and insert new questions. It also supports the creation of a teacher report. At this early stage, all the questions are introduced via this interface, since the automatic question generation module is not yet integrated.

Following the practice session, the system updates the student model, which in this baseline version is a simple word histogram, but will integrate more sophisticated constraints in the near future.

4. Linguistic resources

4.1. Portuguese dictionary

Searching for the meaning of unknown words is a frequent activity for REAP users, which motivated the integration of a Portuguese dictionary. We opted for remote access to an electronic dictionary, from Porto Editora, which displays the meaning, together with the part of speech (POS) tag of each searched (possibly inflected) word. In order to communicate with the dictionary server, an intermediate server/proxy was needed to register the words that are looked up and to update student models. One way to do this is to use *AJAX* technology that allows the communication between web servers using *XML* requests. However, modern browsers do not allow these requests to access outside domains, in order to prevent cross-side scripting. The adopted solution was the use of a proxy server that receives the *XML* requests and establishes a connection to the server of Porto Editora, using an *HTTP* connection. Since the access to the dictionary assumes an *ISO-LATIN1* encoding, some format conversions were adopted.

4.2. Web document corpus

REAP.PT uses WPT 05 as the main document repository ("http://xldb.di.fc.ul.pt/wiki/WPT_05_in_English"). This collection of over ten million documents was retrieved from the Portuguese web obtained by the crawler of the Tumba! search engine, produced by the XLDB Node of Linguateca. The contents were crawled in 2005 and have been harvested among documents written in Portuguese either hosted in a .pt domain or hosted in a .com, .org, .net or .tv domain, and referenced by a hyperlink from, at least, one page hosted in a .pt domain. The average document size is 3000 characters.

4.3. Word list

The Portuguese target word list was designed by the University of Algarve, following similar criteria as the lists made for English (namely, the Academic Word List (AWL) [6]), and French (adopted from the "Echelle Dubois-Buyse" [7]). The list includes currently used, general-purpose, academic level vocabulary. The 3000 words are organised by morphologic families; a catalog of most salient, highly reproducible, different word senses and their corresponding syntactic structures is associated to ambiguous words.

Grade Level	#Books	#Word Tokens	#Word Types
5	5	367,584	18,048
6	6	436,814	21,409
7	6	510,350	25,859
8	5	434,814	21,409
9	7	862,754	31,944
10	8	1,163,924	40,966
11	5	962,800	36,427
12	5	1,085,640	36,229
Total	47	6,862,024	94,857

Table 1: Statistics of the school corpus for each level.

4.4. Documents classified by level

As in other versions of REAP, the standard unit for reading difficulty is the grade level. This first version of REAP.PT is intended both for native high school students and non-native (L2) students. Given that we had no access to enough materials for training distinct level classifiers for the latter, we opted for training classifiers for levels 5-12. The training and test corpora consist of 47 textbooks and exercise books. The statistics are shown in Table 1. One book from each grade contitutes the held-out test set. The same literary texts may be included in more than one text book from the same level.

This test set of textbooks was complemented by a set of national exams ("http://www.gave.min-edu.pt") for the 6th, 9th and 12th levels (5, 7 and 6 exams, respectively).

5. Document filtering

The search module is responsible for retrieving from this web-based corpus the texts satisfying particular pedagogical constraints such as readability level, text length, and containing words from the target list that students should learn. It is also responsible for matching these documents with the student preferences in terms of topic. This filtering stage is a very valuable tool for teachers, saving them time in searching for motivating materials of appropriate quality, readability and topic.

The text length filter removes documents that have less than 200 words or more than 1000 words. Documents containing profanity words (from a list of 160 words) are also removed. In addition, documents containing just word lists are also filtered out. A POS trigram model was computed on a newspaper corpus of 2.6 million words. Given a new document, we computed its POS trigram vector and the cosine similarity between this vector and the model vector. Documents with a score less than a threshold of 0.8 (for example) were discarded.

A pipes-and-filters architecture was adopted to integrate all the filtering stages. The readability and topic classification stages will be described in more detail next.

5.1. Readability

The baseline model is based on lexical features, such as statistics of word unigrams. It can be improved with grammatical features, as in the English version of REAP [8]. Our experiments with Support Vector Machines (SVMs) were made using the SMO tool implemented in WEKA [9]. At this stage, no lemmatization was adopted for this purpose. The use of lemmatization needs further research, as some verbal forms (nowadays

	Correlation	RMSE	Adjacent Acc.
Cross-validation set	0.956	0.676	0.876
Held-out Test set	0.994	0.448	1.000
Exams test set	0.898	1.450	0.550

Table 2: Evaluation of the readability classifier.

mostly found in literary texts and not in normal conversations) may influence reading difficulty.

While grade levels are assigned evenly spaced integers, the ranges of reading difficulty corresponding to these grades are not necessarily evenly spaced. In order to take this into account, [10] tested models for increasing assumptions about the relationships between grade values: nominal, ordinal, interval, and ratio. The best results were obtained using the Proportional Odds (PO) Model [11], which assumes an ordinal relationship. This was the justification for adopting this PO model as well.

The results are shown in Table2, for the 10-fold cross-validation set, the held-out test set and the exams test set. Following [10], we adopted as metrics for evaluating the performance of reading difficulty predictions the root mean square error (RMSE), Pearsons correlation coefficient, and accuracy within 1 grade level. It is interesting to notice that for most exams, the assigned level is either correct or one-level below.

5.2. Topic classification

The current topic classifier was originally developed for broadcast news (BN). It was motivated by the need to compare the performance of our BN topic segmentation and classification modules with the manual topic boundaries and labels done by a professional media watch company [12]. All stories are indexed using 10 topic labels (Economy, Education, Environment, Health, Justice, Meteorology, Politics, Security, Society, and Sports). Classification into multiple classes is very frequent.

REAP.PT currently integrates this 10-class classifier due to the better quality of the trained models, although the set of topics is not yet the target one for teaching purposes. For each of these 10 classes, topic and non-topic unigram language models were created using the stories of the media-watch corpus, which were pre-processed in order to remove function words and lemmatize the remaining ones. Topic classification is based on the log likelihood ratio between the topic likelihood $p(W/T_i)$ and the non-topic likelihood $p(W/\overline{T_i})$. The detection of any topic in a story occurs every time the correspondent score is higher than a predefined threshold. The threshold is different for each topic in order to account for the differences in the modeling quality of the topics. The average accuracy is 91.8% on a held-out test set.

6. Oral comprehension

There is common agreement that one of the most striking differences between Brazilian and European Portuguese varieties concerns vowel reduction, which is much more extreme in the latter ([13], [14]). In the European variety unstressed high vowels are often deleted and rather long consonant clusters may surface within as well as and across word boundaries, which are not allowed in the Brazilian variety. This makes European Portuguese typically more difficult to understand for foreign learners and is one of the motivations for including audio playing options in REAP.PT as a first step towards integrating oral comprehension in the system.

We endeavour to familiarize the student with the way each word/sentence sounds in two ways: by integrating a text-to-speech synthesizer (TTS) for European Portuguese, and by letting the student learn not only from text documents but also from multimedia documents which contain audio (and possibly video) as well.

The first option is available for every text document. Students can highlight words or word sequences as long as they want in the document and click on the "listen" option. When searching for the meaning of a particular word, the dictionary window also includes the same listening option. For this purpose, we have integrated DIXI, a concatenative unit selection synthesizer [15] based on Festival [16].

The second option involves multimedia documents that may consist either of pre-recorded digital talking books (DTB), or broadcast news stories. Digital talking books, also known as audio books, are most often used for entertainment and e-inclusion applications (e.g. for visually impaired or dyslexic users). Their use in the area of CALL is not so typical [17], but the possibility of listening to the audio signal of word or word sequence in the text may be very important for L2 students. Moreover, compared with other languages, such as English, there are relatively very few materials for learning European Portuguese that include controlled quality recordings both at the segmental and prosodic levels. So DTBs could be a very good way to become familiar with the language, for non-native students of Portuguese.

The alignment of each spoken word with the read text is achieved using our automatic speech recognition system (ASR) in a forced alignment mode. AUDIMUS [18] is a hybrid recognizer whose acoustic models combine the temporal modeling capabilities of hidden Markov models with the pattern classification capabilities of multi-layer perceptrons. Its decoder, which is based on weighted finite state transducers, proved very robust even for aligning very long recordings (a 2-hour long book could be aligned in much less than real-time).

The repository of aligned DTBs is still quite limited, being mostly used for demonstration purposes, but it already includes a wide range of genres: fiction, poetry, childrens stories, and didactic text books. This repository, however, does not typically cover the potentially very wide areas of interest of L2 students. That was the main motivation for adding a totally diferent repository of BN stories, taking advantage of the large corpus that was manually transcribed for the purpose of training/testing AUDIMUS. The corpus includes over 80h of manually transcribed news shows. Because transcriptions were only manually aligned at the utterance level, AUDIMUS is again used in its forced alignment mode to produce word-level alignment.

Another possibility that we are currently investigating is the use of automatically transcribed BN stories, which would have the obvious advantage of allowing the student to learn from very recent documents. Here the text is the output of AUDIMUS, enriched with punctuation and capitalization. Each BN show is split into its constituent stories and each story is topic classified using the algorithm described above. In spite of the fact that both lexical and language models are dynamically adapted each day with the latest news retrieved from on-line newspapers, as required for on-line subtitling [18], the word error rate still exceeds 20% in spontaneous speech segments or very noisy ones. For the segments spoken by the anchor who typically introduces each news story, the error rate typically does not exceed 10%.

In order to investigate whether this error rate is sufficiently

low for CALL purposes, the current interface uses a different font color for every word that was recognized with a confidence level below a given threshold (82%). We are currently exploring additional readability filters that take into account the global confidence level of a transcript segment (automatically detected as belonging to a given speaker), and only allow contiguous transcript segments without low-confidence parts, as the ones that typically constitute the introduction of a news story by the anchor. The impact of the recognition errors on the different filters is also worth investigating.

7. Conclusions and future work

REAP is a tutoring platform that may integrate a large number of language processing tools and resources. Despite the enormous amount of work that still needs to be done to enhance the Portuguese version, we have already built the minimum requirements for progressing to the first field trials, scheduled for June 2009, in the University of Algarve.

REAP has also been ported to a prototype French version. Porting to Portuguese has extended our experience on the general issues encountered when porting this software to other languages. Apart from encoding issues, an obvious difficulty is the relative lack of computational linguistic resources in languages different from English. For example, when building a readability classifier, a stemmer might be needed. In addition, in order to use syntactic features in the classifier, a syntactic parser is required. Although lexical features might be enough, it could be argued that in some languages, syntactic features are more important, for example for languages that are morphologically richer than English. If a set of web pages is not previously available, one may need to generate queries to crawl the web, which may mean applying a morphological generator to the focus words for query expansion. Such a tool was harder to find for French, and the same can be probably said for many other languages. A POS tagger may be used to measure text quality. Again, there were no directly available tools for French, which motivated their adaptation from English. The training materials for the readability and topic classifiers are two other very important resources not always easy to find. Dictionary integration may be a major issue because open source dictionaries such as wikicionary do not provide enough quality. Moreover good dictionaries are not available in electronic format or have access restrictions. Finally, in order to generate synonym questions, one might prefer building a thesaurus automatically rather than rely on non free resources such as EuroWordNet.

REAP's flexibility has been improved by adding audio playing capabilities, based on either text-to-speech synthesis or automatic alignment of previous recorded documents (DTBs and BN stories). Their impact for L2 learners of European Portuguese is one of the target goals of the forthcoming tests.

8. Acknowledgments

The authors would like to thank Lisboa Editora and Porto Editora for giving us access to the to the training corpus and the on-line dictionary, and their colleagues Hugo Meinedo and Luís Figueira for their help. This work was supported by project CMU-PT/HuMach/0053/2008, sponsored by FCT.

9. References

[1] M. Heilman, K. Collins-Thompson, J. Callan, and M. Eskenazi, "Classroom success of an intelligent tutoring system for lexical practice and reading comprehension," in *Proc. Interspeech 2006*, Philadelphia, Sep. 2006.

[2] L. Amaral, V. V. Metcalf, and D. Meurers, "Language awareness through re-use of nlp technology," in *Preconference Workshop on NLP in CALL Computational and Linguistic Challenges. CALICO*, Manoa, Hawai, USA, 2006.

[3] E. Miltsakaki and A. Troutt, "Read-x: Automatic evaluation of reading difficulty of web text," in *Proc. E-Learn 2007*, Quebec, Canada, 2007.

[4] E. Miltsakaki, "Matching readers' preferences and reading skills with appropriate web texts," in *Proc. EACL 2009 Demo session*, Athens, Greece, 2009.

[5] K. Sheehan, I. Kostin, and Y. Futagi, "Sourcefinder: A construct-driven approach for locating appropriately targeted reading comprehension source texts," in *Proc. of the SLaTE Workshop on Speech and Language Technology in Education*, Farmington, Pennsylvania USA, Oct. 2007.

[6] A. Coxhead, "A new academic word list," *TESOL Quarterly*, vol. 34, no. 2, pp. 213–238, 2000.

[7] F. Ters, G. Mayer, and D. Reichenbach, *L'Echelle Dubois-Buyse*. Editions M.D.I, 1995.

[8] K. Collins-Thompson and J. Callan, "A language modeling approach to predicting reading difficulty," in *Proceedings of the HLT/NAACL 2004 Conference*, Boston, 2004.

[9] I. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques*. San Francisco: Morgan Kaufmann, 2005, 2nd edition.

[10] M. Heilman, K. Collins-Thompson, and M. Eskenazi, "An analysis of statistical models and features for reading difficulty prediction," in *Proc. 3rd Workshop on Innovative Use of NLP for Building Educational Applications, Association for Computational Linguistics*, Columbus, Ohio, USA, Jun. 2008.

[11] P. McCullagh, "Regression models for ordinal data," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 42, no. 2, pp. 109–142, 1980.

[12] R. Amaral and I. Trancoso, "Topic segmentation and indexation in a media watch system," in *Proc. Interspeech '2008*, Brisbane, Australia, Sep. 2008.

[13] M. H. Mateus and E. d'Andrade, *The Phonology of Portuguese*. Oxford: Oxford University Press, 2000.

[14] P. Barbosa and E. Albano, "Brazilian Portuguese - Illustrations of the IPA," *Journal of the Int. Phonetic Association*, vol. 34, no. 2, 2004.

[15] S. Paulo, L. Oliveira, C. Mendes, L. Figueira, R. Cassaca, C. Viana, and H. Moniz, "DIXI - a generic text-to-speech system for European Portuguese," in *PROPOR'2008 - 8th International Workshop on Computational Processing of the Portuguese Language*. Curia, Portugal: LNAI 5190, Springer-Verlag, Sep. 2008.

[16] A. Black, P. Taylor, and R. Caley, "The Festival speech synthesis system," Dec. 2002.

[17] I. Trancoso, A. Serralheiro, C. Viana, D. Caseiro, and M. I. Mascarenhas, "Digital talking books in multiple languages and varieties," in *3rd Language & Technology Conference*, Poznan, Poland, Oct. 2007.

[18] H. Meinedo, M. Viveiros, and J. Neto, "Evaluation of a live broadcast news subtitling system for Portuguese," in *Proc. Interspeech 2008*, Brisbane, Australia, Sep. 2008.