MULTIMEDIA LEARNING MATERIALS

J. Lopes[1 2], *I. Trancoso*[1 2], *R. Correia*[1 2], *T. Pellegrini*[1], *H. Meinedo*[1], *N. Mamede*[1 2], *M. Eskenazi*[3]

[1] INESC-ID Lisbos
[2] Instituto Superior Técnico, Lisboa, Portugal
[3] Carnegie Mellon University, Pittsburgh, USA

ABSTRACT

This paper describes the integration of multimedia documents in the Portuguese version of REAP, a tutoring system for vocabulary learning. The documents result from the pipeline processing of Broadcast News videos that automatically segments the audio files, transcribes them, adds punctuation and capitalization, and breaks them into stories classified by topics. The integration of these materials in REAP was done in a way that tries to decrease the impact of potential errors of the automatic chain in the learning process.

Index Terms— CALL, broadcast news transcription

1. INTRODUCTION

The Broadcast News (BN) processing system developed at the Spoken Language Systems Lab of INESC-ID integrates several core technologies, in a pipeline architecture: jingle detection, audio segmentation, automatic speech recognition, punctuation, capitalization, topic segmentation/indexation, and summarization. Language identification and translation modules were recently added. The first modules of this system were optimized for on-line performance, given their deployment in the fully automatic subtitling system that is running on the main news shows of the public TV channel in Portugal, since March 2008. Besides this on-line use, the system has multiple off-line applications, integrated in media watch systems, video search engines, etc. [1]. The most recent application of this BN processing system is a source of multimedia learning materials.

This paper concerns the integration of these materials in the Portuguese version of REAP, 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 [2].

REAP provides a very flexible platform, where 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.

REAP.PT [3], the Portuguese version of REAP, integrates several innovations, in an effort to familiarize students with the way each word/sentence sounds in this language, for which the availability of audio materials and corresponding transcriptions is relatively scarce, and the typical vowel reduction makes it hard to understand for non-native speakers. The first step was the integration of a text-to-speech synthesizer for European Portuguese. Students can select any text and listen to the corresponding synthesis. 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 [4] based on Festival [5].

The second step was the integration of audio books, also known as digital talking books (DTBs). The alignment of each spoken word with the read text is achieved using our automatic speech recognition system in a forced alignment mode. Although the main applications of audio books are in e-inclusion and entertainment areas, we believe that their use in the area of CALL (Computer Assisted Language Learning) has an unexplored potential, giving L2 students the possibility of listening to the audio signal of isolated words or word sequences, frequently spoken by professional speakers, with very natural articulation and intonation. Audio books, however, typically consist of full stories, much longer than the reading sessions recommended in REAP. In European Portuguese, moreover, the repository is much too short to provide students with a large choice of topics, and they are not typically very recent documents, as desired.

This was the motivation for including BN materials instead, which besides being very recent, include a wide choice of short stories on different topics, and have the added value of video. All the processing, however, is fully automatic, and students should be alerted to the eventual presence of errors.

This paper is structured into two main sections: one on the BN pipeline system, discriminating its main modules and characterizing their performance, and another on the integration of this BN processing system within REAP.PT.

2. THE BROADCAST NEWS PIPELINE

The first stage of the BN pipeline is a jingle detection module that may be tuned to different TV channels, detecting the start and end of the news show, and excluding publicity segments. It is based on Multi-Layer Perceptrons (MLPs). All the subsequent modules contribute information to an XML file.

2.1. Audio segmentation

The second stage is the audio segmentation (or diarization) module which includes several components: three for classification (Speech / Non-speech, Gender, and Background), one for speaker clustering, one for acoustic change detection, and a speaker identification module that identifies very frequent speakers for which models have been previously built, such as anchors. All classifiers make extensive use of MLPs. Although audio segmentation errors may impact on the performance of the subsequent modules, their direct impact in the REAP.PT framework is comparatively negligible, as the system only directly uses the speaker clustering information to break what otherwise would be very long paragraphs into speakers turns. The clusters are numbered according to gender, male speaker clusters starting in number 1000, female speakers with 2000, and child speakers with 3000. Clustering errors are frequent in overlapping segments, or artifical collages with no pauses between them. Also possible is the existence of several clusters for the same speaker, typically characterized by different background counditions.

2.2. Automatic speech recognition

The third stage of the pipeline is speech recognition. Our automatic speech recognition engine named Audimus [6] is a hybrid system that combines the temporal modeling capabilities of Hidden Markov Models (HMMs) with the pattern discriminative classification capabilities of MLPs. The MLP/HMM acoustic model combines posterior phone probabilities generated by three phonetic classification branches, based on PLP (Perceptual Linear Prediction), Log-RASTA (log-RelAtive SpecTrAl), and MSG (Modulation Spectrogram) features. The decoder is based on the Weighted Finite-State Transducer approach, where the search space is a large transducer that results from the integration of the HMM/MLP topology transducer, the lexicon transducer and the language model one. Besides the recognized words, the decoder outputs a word confidence measure, computed by combining several features through a maximum entropy classifier, whose output represents the probability of the word being correct.

The initial acoustic model was trained with 46 hours of manually annotated BN data collected from the public Portuguese TV. Unsupervised training has been adopted at a second stage, using as training data all the words that were recognized with a confidence measure above 91.5%. The first iteration used 378 hours of useful training speech data, 332 of which were automatically annotated. The second iteration used a total of 1000 hours of data mostly news shows from several TV channels. The MLPs model 38 three state monophones plus a single-state non-speech model (silence), and 385 phone transition units which were chosen to cover a very significant part of all the intra-word transition units present in the training data.

One of the main characteristics of this recognizer is the dynamic adaptation of language and lexical models which is performed daily. The Language Model (LM) is a statistic 4-gram model and results from the interpolation of three specific LMs: a backoff 4-gram LM, trained on a 700M word corpus of newspaper texts; a backoff 3-gram LM estimated on BN transcripts; and a backoff 4-gram LM estimated on the web newspapers texts collected from the previous seven days. The final interpolated language model is a 4-gram LM, with Kneser-Ney modified smoothing.

The vocabulary is also adapted on a daily basis from web newspaper texts [7]. This daily modification implies a re-estimation of the LM and retraining of the word confidence measure classifier. After the 100k word vocabulary selection, the pronunciation lexicon is built automatically by dividing the words into two categories. The ones for which we are able to produce a correct pronunciation using either an in-house lexicon or a rule-based grapheme-to-phone (GtoP) conversion module [8], and the ones that do not follow the Portuguese pronunciation rules, such as spelled acronyms (that require a special set of rules) and foreign words. The latter are further subdivided, in order to identify the ones that exist in the public domain lexicon provided by Carnegie Mellon University, for which nativized pronunciations are derived by rule. For the words not included in this lexicon, grapheme nativization rules are applied prior to using the GtoP module to generate the pronunciation. The final multiple-pronunciation lexicon generally includes 114k entries.

The Word Error Rate (WER) for our 2007 evaluation set, RTP07, composed by six one-hour news shows, was 18.4%, using a fixed LM/vocabulary. As in most recognition systems, the performance significantly differs from clean read speech conditions (typically below 10%) to spontaneous speech or very noisy environments (typically above 20%).

2.3. Punctuation and capitalization

The fourth module is in charge of applying punctuation and capitalization [9] to the raw output provided by the recognizer. Both approaches are based on the Maximum Entropy method. Capitalization explores only lexical cues. Results for automatically transcribed BN (F-measure=75.4%) were as expected worse than for manually transcribed BN (F-measure=85.6%), thus showing the effects of recognition errors.

Punctuation explores lexical, acoustic and prosodic cues. The earlier approach targeted only commas and full stops.

Typical evaluation metrics for these two tasks are F-measure, and Slot Error Rate (SER). Results for full stop detection (F-measure 69.7%) were better than for comma detection (50.9%). Research on the detection of question marks is still at a very early stage (F-measure=27.3%) and therefore was not yet integrated.

The performance of both modules seems quite acceptable for read speech segments. The punctuation and capitalization module also frequently integrates a normalization stage that deals with the typical way numerical entities should be presented. This module is the last one that is integrated in the on-line subtitling system. All subsequent modules are off-line.

2.4. Topic segmentation and indexation

The goal of topic segmentation module is to split the BN show into the constituent stories. This is done taking into account the characteristic structure of broadcast news shows [10]. They typically consist of a sequence of segments that can either be stories or fillers (i.e. headlines / teasers marked by the audio segmentation module). The fact that all stories start with a segment spoken by the anchor, and are typically further developed by out-of-studio reports and/or interviews is the most important heuristic that can be exploited in this context. Hence, the simplest segmentation algorithm is the one that starts by defining potential story boundaries in every transition non-anchor / anchor. This heuristic has been refined in a CART-based approach (Classification and Regression Trees) that deals with double-anchor shows, the presence of local commentators, or a thematic anchor for particular sections (such as sports). This module obtained an F-measure of 84%. As expected, its performance is strongly dependent on the audio segmentation module.

The topic indexation module assigns one or multiple topics to each story, out of a set of 10[11]: Economy, Education, Environment, Health, Justice, Meteorology, Politics, Security, Society, and Sports. A further classification into National and International is also done. For each of these classes, topic and non-topic unigram language models were created using the stories of a 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 and the non-topic likelihood. 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.

Although summarization is a current research topic in the group, the REAP system adopts the simplest strategy of using the first sentence as a summary of each BN story. This extractive summarization approach is particularly suited to BN news.

3. INTEGRATION IN REAP.PT

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 words from a 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, multimedia documents, and topic interests. The first two are text-based and 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 (as shown in the previous section) 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.

The BN module is obviously part of the multimedia section. An example of the interface is shown in figure 1. The left part shows the topic selection menu, displaying a chronologically ordered sequence of stories for each topic. By clicking on a story, the student can watch the corresponding video with subtitles on the left side, whereas the right side shows the transcribed text, structured by speaker turns, for improved readability.

3.1. Document filtering

This transcribed text shares many features with the text-based documents, but there are major differences. The first one is on the filtering stage which is used to retrieve adequate documents from the web-based corpus on the topics of preference of each student. The texts have to satisfy particular pedagogical constraints such as text length (above 300 words), and they need to contain a minimum number of words from the target list that students should learn. Text documents containing profanity words (from a list of 160 words) are also removed. In addition, documents containing just word lists are also filtered out.

Most of these filters are bypassed on the multimedia documents. The text length filter was discarded, as the average number of words per story is 280 words in our evaluation set. Topic segmentation module does not allow too short stories, merging them with neighbouring stories, to avoid the risk of assigning topics on the basis of too little material. In exceptional cases, such as major disasters, stories may be very long, and such as for text documents, a warning to stop reading after 1000 words is given to the student. The profanity filter was also discarded, as these words are excluded at the ouput of the recognizer. Likewise, word lists would not be typically produced by the language models, so this filter was also discarded.

The last two filtering stages (topic and readability) are almost identical in the text-based and multimedia documents,

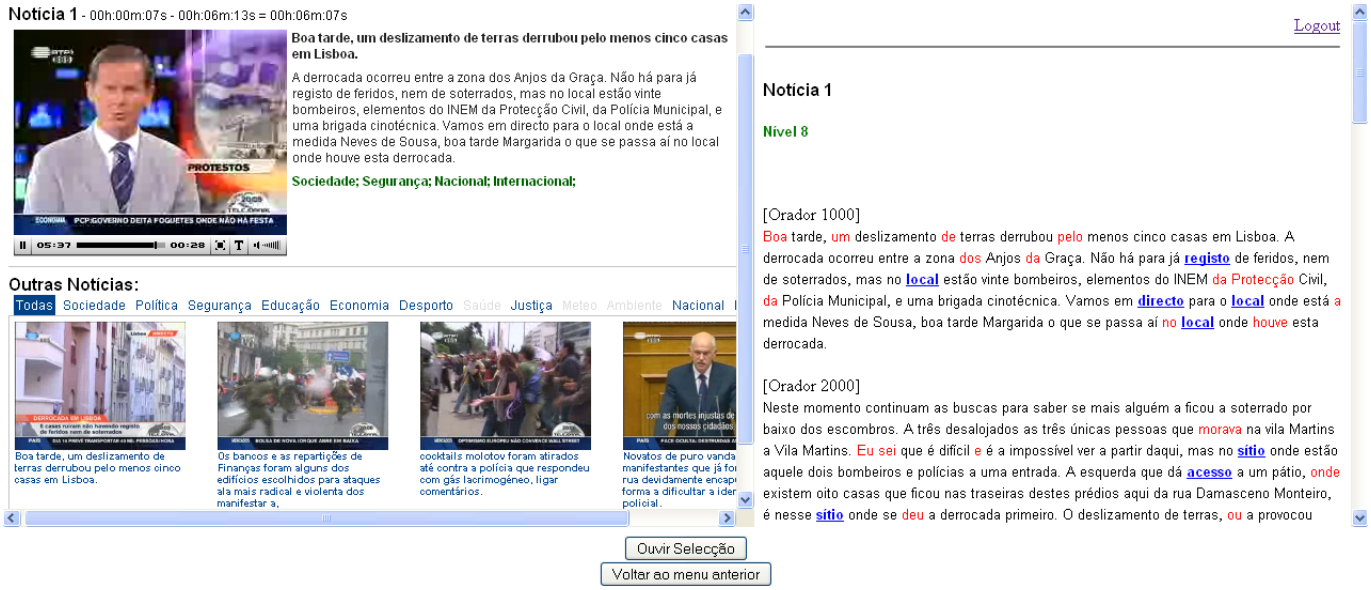

Fig. 1. Example of the BN interface in REAP.PT.

the main difference being that recognized words with a confidence measure below a given threshold are not taken into account.

Due to the difficulties in collecting electronic materials for non-native students classified by level, the readability classifier was trained on the basis of textbooks for native Portuguese students, from grades 5-12, a strategy that was also followed in other versions of REAP. The initial model is based on lexical features, such as statistics of word unigrams. Our experiments with Support Vector Machines (SVMs) were made using the SMO tool implemented in WEKA [12]. At this stage, no lemmatization was adopted for this purpose. The use of lemmatization needs further research, as some verbal forms (nowadays mostly found in literary texts and not in normal conversations) may influence reading difficulty.

The readability classifier already produced interesting results on a held-out test set of textbooks, when evaluated according to typical metrics of reading difficulty predictions [13]: root mean square error (0.448), Pearsons correlation coefficient (0.994), and accuracy within 1 grade level (1.000).

It is interesting to notice that the news stories in our BN evaluation set were classified between levels 7 and 11, with an average of 8. In this classification, however, there are many words that are not taken into account, as the vocabulary covered by the training set of textbooks, although very large (close to 95k words), is very different from the dynamic recognition vocabulary. More sophisticated readability models are being investigated [14] [15] [16].

3.2. Document display

The XML file produced by the BN processing chain serves as input to the transcribed text of each news displayed in the right part of the screen.

All the words in the transcribed text that belong to the target word list the student is supposed to learn at each stage are underlined, just as for text documents. This target word list, which henceforth will be designated as P-AWL (Portuguese Academic Word List [17]), is composed by 2,019 different *lemmas*, together with their most common inflections, each one tagged with the correspondent part-of-speech (POS) , totallizing 33,284 words. Similarly to the English and French counterparts, it is a vocabulary specially designed for language learning.

Students can also search for the meaning of any unknown words in the original or automatically transcribed texts, 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. Clicking on the audio button in the dictionary pop-up triggers the synthesizer. However, clicking on "listen to selection" of a sequence of words in the transcribed text triggers the audio player of the recorded BN segment, instead.

The other main difference between original texts and automatically transcribed texts is the occurrence of recognition errors. In order to alert students to their potential presence, words recognized with a low confidence (below 82%) are shown in red. The relevance of confidence measures for selecting learning materials has in fact been a major motivation

for our recent work on error detection [18]. This work explored new features gathered from other knowledge sources than the decoder itself: a binary feature that compares outputs from two different ASR systems (word by word), a feature based on the number of hits of the hypothesized bigrams, obtained by queries entered into a very popular Web search engine, and finally a feature related to automatically infered topics at sentence and word levels. The combination of the baseline decoder-based features and the first two additional features led to significant improvements in detecting recognition errors, from 13.87% to 12.16% classification error rate, comparing to a baseline using only decoder-based features.

When faced with speech from other languages, a relatively frequent occurrence in Portuguese BN, the word confidence module typically outputs a stream of very low confidence words. Although this might constitute enough warning for a non-native student not to use these materials for learning, a cleaner approach would be to mark the segment as "not Portuguese". For this purpose, a language verification module that is not part of the on-line processing chain has been integrated in the system, providing more information to the XML file.

The extension of word-based confidence measures to sentence-based is also interesting from a CALL point of view. Given the much lower confidence values that are typically found in spontaneous speech, such segments may almost be identified by their red color, but an explicit marking of "spontaneous" may turn out as useful as the language attribute, and is therefore also in our plans.

Every action by the student is tracked during the reading session, namely the access to the dictionary, in order to keep updating his/her progress. The reading session, whether it has been based on text documents or multimedia documents, is followed by a series of cloze, or fill-in-the-blank, questions about the words that were highlighted. For the time being, a set of 6000 cloze questions was manually selected. However, the set of distractors was automatically generated. In fact, a recent study compared different procedures for this generation: manual, random, graphemic, and phonetic [19].

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.

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. CONCLUSIONS AND FUTURE WORK

This paper described the inclusion of audio playing options in REAP.PT, namely thorugh the integration of a synthesizer, the use of audio books, and the integration of automatically processed BN materials, which was the focus of the current work.

The first field trials of REAP.PT were delayed for logistic reasons, and are scheduled for the next semester in the University of Algarve. Despite the absence of a formal evaluation, the multimedia BN interface for REAP.PT has been informally tested by more than 10 non-native students, who found it very motivating, and have already provided useful feedback. They unanimously said that highlighting each word as it is spoken in the right part of the screen would be very useful, so this is a feature we must immediately integrate.

They also agreed that the possibility of slowing down the audio playback might be very useful, although the quality degradation incurred in very basic speech rate change methods may be counter productive. In fact, none of the typical strategies compensates the effects of vowel reduction, a phenomenon that is much more extreme in European Portuguese than in Brazilian Portuguese ([20]). In fact, 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.

The inclusion of multimedia documents in REAP opens up many possibilities. We are currently engaged in designing multimedia word games to make the reading sessions even more appealing. The possibility of using automatically transcribed TED talks is also a contribution in the same direction.

The extension of REAP.PT to other varieties of Portuguese is straightforward. Our recognition system needed to be ported to both Brazilian and African varieties [21], as specially in the first case, the degradation obtained with a recognizer trained for European Portuguese was unacceptable (WER above 50% for BN). However, most of the other modules in the pipeline system did not need retraining.

REAP is a tutoring platform that in its original version already integrated a large number of natural language processing tools. Despite the enormous amount of work that still needs to be done to enhance the Portuguese version, we have shown that it may also integrate a very large number of spoken language processing tools, to make it more appealing to language students.

5. ACKNOWLEDGEMENTS

This work was funded by FCT projects PTDC/PLP/72404/2006 and CMU-PT/HuMach/
0053/2008. INESC-ID Lisboa had support from the POSI Program of the "Quadro Comunitário de Apoio III".

6. REFERENCES

[1] M. Ostendorf, B. Favre, R. Grishman, D. Hakkani-Tur, M. Harper, D. Hillard, J. Hirschberg, H. Ji, J. Kahn, Y. Liu, E. Matusov, H. Ney, A. Rosenberg, E. Shriberg, W. Wang, and C. Wooters, "Speech segmentation and spoken document processing," *IEEE Signal Processing Magazine*, vol. 25, no. 3, pp. 59–69, 2008.

[2] 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, Sept. 2006.

[3] L. Marujo, J. Lopes, N. Mamede, I. Trancoso, J. Pino, M. Eskenazy, J. Baptista, and C. Viana, "Porting REAP to European Portuguese," in *Proc. SLATE 2009*, UK, 2009.

[4] 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, Sept. 2008, LNAI 5190, Springer-Verlag.

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

[6] J. Neto, H. Meinedo, M. Viveiros, R. Cassaca, C. Martins, and D. Caseiro, "Broadcast news subtitling system in Portuguese," in *Proc. ICASSP'2008*, Las Vegas, Mar 2008.

[7] C. Martins, A. Teixeira, and J. Neto, "Dynamic language modeling for a daily broadcast news transcription sys tem," in *Proc. ASRU 2007*, Kyoto, Japan, 2007.

[8] D. Caseiro, I. Trancoso, L. Oliveira, and C. Viana, "Grapheme-to-phone using finite state transducers," in *Proc. 2002 IEEE Workshop on Speech Synthesis*, Santa Monica, CA, USA, Sept. 2002.

[9] F. Batista, D. Caseiro, N. Mamede, and I. Trancoso, "Recovering capitalization and punctuation marks for automatic speech recognition: Case study for the Portuguese broadcast news," *Speech Communication*, vol. 50, no. 10, pp. 847–862, Oct. 20078.

[10] R. Barzilay, M. Collins, J. Hirschberg, and S. Whittaker, "The rules behind roles: Identifying speaker role in radio broadcast," in *Proc. AAAI 2000*, Austin, USA, July 2000.

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

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

[13] 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, June 2008.

[14] P. Kidwell, G. Lebanon, and K. Collins-Thompson, "Statistical estimation of word acquisition with application to readability prediction," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*. Association for Computational Linguistics, 2009.

[15] J. Callan and M. Eskenazi, "Combining lexical and grammatical features to improve readability measures for first and second language texts," in *Proceedings of NAACL HLT*, 2007, pp. 460–467.

[16] S.E. Schwarm and M. Ostendorf, "Reading level assessment using support vector machines and statistical language models," in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, 2005.

[17] J. Baptista, N. Costa, J. Guerra, M. Zampieri, M. Cabral, and N. Mamede, "P-AWL: Academic Word List for Portuguese," in *Proc. PROPOR 2010, LNAI 6001*, 2010.

[18] T. Pellegrini and I. Trancoso, "Improving ASR error detection with non-decoder based features," in *accepted for Interspeech 2010*, Makuhari, Japan, 2010.

[19] R. Correia, J. Baptista, N. Mamede, I. Trancoso, and M. Eskenazi, "Automatic generation of cloze question distractors," in *Accepted for publication in Proc. Workshop on Second Language Studies: Acquisition, Learning, Education and Technology*, Tokyo, Japan, Sept. 2010.

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

[21] A. Abad, I. Trancoso, N. Neto, and C. Viana, "Porting an European Portuguese broadcast news recognition system to Brazilian portuguese," in *Proc. Interspeech 2009*, Brighton, UK, 2009.