Listening Comprehension Games for Portuguese: Exploring the Best Features

Rui Correia[1,2,3], *Thomas Pellegrini*[1], *Maxine Eskenazi*[3],
Isabel Trancoso[1,2], *Jorge Baptista*[1,4], *Nuno Mamede*[1,2]

[1]INESC-ID Lisboa, Portugal
[2]IST, Lisboa, Portugal
[3]Language Technologies Institute, Carnegie Mellon University, USA
[4]University of Algarve, Portugal

{Rui.Correia,Thomas.Pellegrini}@inesc-id.pt

Abstract

This paper investigates which features would be the best to include in a listening comprehension game for European Portuguese. The goal is to develop a method to motivate non-native speakers of Portuguese to engage themselves in activities intended to develop their listening comprehension skills. The approach adopted herein consisted in providing students with several sets of exercises, each set with some slight variations over the previous one. Finally, a questionnaire was presented to the students with the intent of receiving feedback about the most adequate functionalities. The paper also explores students main difficulties with the language so as to define learning models able to adapt the difficulty level of the exercises to each individual's language proficiency level.

Index Terms: CALL, Portuguese, Listening Comprehension, Games

1. Introduction

The traditional view of the student as a sufficiently self-motivated individual to engage in learning through book reading, attending classes and doing homework, all by himself, has changed over the past few years. In a globally, web-connected World, with permanent exposure to interactive applications and sensory-appealing gadgets, it is increasingly difficult to keep students motivated in their academic subject matters.

The CALL area (Computer Assisted Language Learning) tries to fill that gap by bringing together technology and language education, in order to provide 'easy to use', interactive and exciting language learning oriented systems.

REAP.PT [1] (REAding Practice for PorTuguese) is one of these systems. As the name states, REAP.PT is a browser based tutoring system that has reading activities as a learning methodology. The main concept behind this system is to provide students with real texts, collected from the Web (ClueWeb09[1]), that are recent and that actually match their personal interests. Hence typically different students will have different interactions with the system, since the latter retrieves the learning materials matching the student's level and interests.

Expanding REAP.PT by introducing a listening comprehension module results from two major factors. On the one hand, Portuguese has one of the richest phonology among the Latin languages, which turns listening comprehension skills hard to master. For the European Portuguese variety (EP) in particular, the difficulties are even stronger since it is characterized by strong vowel reduction. On the other hand, students learning EP do not have easy access to materials in this variety. Our in-house repository of EP broadcast news, which have been daily stored and automatically transcribed since 2009, may contribute to filling this gap.

Broadcast news (BN) meet the requirements of REAP.PT, containing very recent material, segmented into short stories that are automatically classified with topics (such as sports, economy, etc.). Previous experiments with BN data [2] showed that the language level of the stories span over the 7^{th} and the 11^{th} grades of native European Portuguese students (with an average corresponding to the 8^{th} grade). One should notice that this classification was based on complete stories and not only on single sentences. However, finding stories adequate to each student level is not the topic of the current paper. Instead, we focus on discovering the best features for building listening comprehension exercises, while leaving the issue of level suitability for future work.

This paper is organized as follows: Section 2 presents some theoretical background and illustrates it with some examples of educational games. Section 3 consitutes the core of the work, presenting the experimental setup and in Section 4 results are analyzed.

2. Related Work

In [3], Richards provides some theoretical background relevant to the listening comprehension process and summarizes the core aspects involved in teaching this particular skill.

There is a clear distinction between written discourse and spoken language: the first one uses as atomic unit the sentence while the latter a single clause. Conversational discourse is also characterized by the need to express meaning efficiently. This may lead to omission of words that are less crucial, disappearance of word boundaries, omission of specific sounds or even substitutions. When facing naturally occurring, spontaneous speech, there is also insertion of sounds that do not contribute to the meaning. Disfluencies can constitute up to 50% of speaking time. Another important factor is the so-called rate of delivery. This is either defined by the pauses the speaker uses between clauses and the actual speech speed. Finally, Richards points out the interactivity of spoken language: gestures, movement, gaze, and facial expressions can express meaning and define the tone of the speech.

Secules et al. [4] showed how listening comprehension skills improve when using the video-based contents on French students. Brett [5] showed that authentic video materials along

[1] http://boston.lti.cs.cmu.edu/Data/clueweb09

with a subtitling feature can increase the students' motivation to engage in these types of tasks.

Recently, games have gained strong interest in the CALL community to support L2 acquisition. These games are referred to as *serious games*, with an educational goal that goes beyond mere entertainment [6].

Examples of such games are Mingoville[2] for children, or Rainbow Rummy [7] and Polyglot Cubed [8] for adult learners.

3. Experimental Setup

In order to understand which features are best for a listening comprehension game, one developed a test session where the users were guided through a set of 18 exercises, ending with a questionnaire designed to elicit the students' preferences. Since the final goal was to integrate the resulting game in the REAP.PT system, the experiment was developed to be accessible online, via a Web browser.

The session was divided into 6 sets of 3 exercises. Each set contained slight variations compared to the previous one. Each exercise consisted of listening to a sentence by either using audio only or using audio and video together. The goal of the exercise also varies along the sets: for the exercises in sets 1 to 3, the student is required to reconstruct the sentence by ordering all of its words and, for the exercises in sets 4 to 6 the student should identify and select only specific words that were present in the sentence. The list of candidate words of all the sets includes both correct words and distractors. Users can play the audio or the video an unbound number of times. It is important to remark that users never have to write down any word. This strategy allows them to focus on listening skills exclusively, so that most spelling skills are not involved in the exercise.

For each exercise, users answered two questions:

- *Which were the main difficulties of the current exercise?* – the user could point out if the sentence was too long, if speech was too fast, if there were unknown words or type in any other difficulty.
- *Are there any errors in the answer, compared to what you heard?* – when in presence of automatically generated exercises, it is essential not to mislead the user with incorrect content. This question aims at finding the sensibility of the users regarding possible recognition errors.

Table 1 summarizes the main differences between the sets, that will be detailed in the following subsections.

Table 1: *Functionalities of the different sets.*

	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6
Video	×	×	✓	✓	✓	✓
Slow down	×	✓	✓	×	×	×
Memorizing	×	×	×	✓	×	×
Recent material	×	×	×	×	×	✓
Karaoke	×	×	×	×	×	✓

3.1. Sets 1–3

Sets 1–3 take the form of a "word puzzle" minigame, in which the student first listens to a sentence, then attempts to form the sentence by selecting the appropriate words from a list of candidates including both the correct words and some distractors. Figure 1 shows an instance of the main interface.

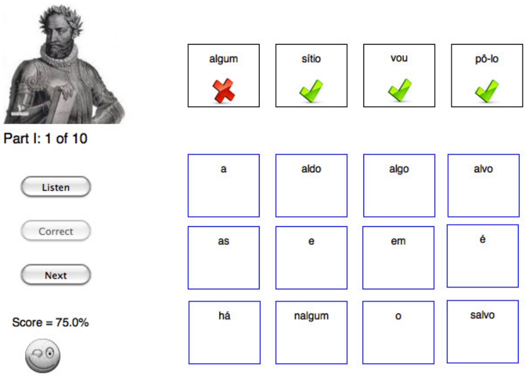

Figure 1: *Interface for the utterance "nalgum sítio vou pô-lo" from Set 1.*

Boxes, each labeled with a single word, can be dragged and dropped one by one with the mouse, into a sequence of empty target boxes. When the student finishes forming the sentence, a visual correction is given by adding a green validation mark or a red cross for each word, when the word is correct or wrong, respectively. Misplaced but correct words are marked with a green background color. A score in percentage appears on the screen. After this correction, the user can still move the words to correct the answer, but the score remains unchanged.

Nine sentences were manually selected from the ALERT corpus [9] – a large set of BN shows, that were transcribed manually. Only clear audio utterances, consisting of well-formed, short sentences with 4 to 10 words were considered.

An Automatic Speech Recognizer (ASR) engine, Audimus [10]), was run over these sentences in order to use as distractors the words that were in competition during the decoding process. These words are phonetically close to the correct words in the target sentence, so they may be considered to constitute good distractors for the exercise. On average, the number of distractors/number of correct words ratio was 2.7, that is, almost three distractors per correct word are presented. This automatic method effectively reduces the need for teacher supervision.

As shown in Table 1, sets 1 and 2 provide only audio, whereas set 3 provides video. Furthermore, set 2 introduces an extra *slow down* feature, allowing the user to listen to the original utterance and, if needed, to the utterance with a scaled speed of 0.8. This feature simulates a lower speech rate and it is expected to help understanding the utterance better. In fact, speaker delivery rate is pointed out as an important aspect in listening comprehension [3]. Finally, set 3 provides the same features as set 2, but with the corresponding BN videos.

3.2. Sets 4–6

The goal of these exercises is to select, among a list of words, the ones the student heard on the clip, instead of having to order all the words to form a sentence. This group of sets mainly differs from the previously described sets in the way utterances are selected. Instead of being guided through a fixed set of exercises, the student has now the opportunity to search for any sequence of words, and then get BN video passages where the query appears. To enable this search feature, the so-called ASR *transcript segments* of BN shows, dating from January 2009 through March 2011, were indexed. Transcript segments, as Richards [3] describes, are speech segments delimited by sig-

[2] www.mingoville.com

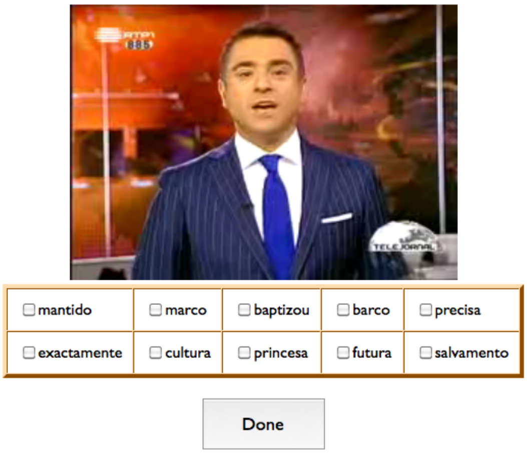

Figure 2: *Interface for the search "barco" (boat) for Set 5.*

nificant pauses, and for that reason, may be viewed as an approximation to clauses.

Figure 2 shows an instance of the main interface of the sets 4–6. These so-called *target words*, are outputted by the ASR engine, and are selected based on a minimum *confidence measure* (CM) of 90%. Confidence measures define how reliable a hypothesized word is. The 100 top unigrams of the BN transcriptions were discarded to ignore very common words, such as articles, pronouns, common collocations, among others.

Moreover, the *transcript segments* were not all indexed. Only segments with less than 15 words and more than 5 target words were selected. Segments with an average CM lower than 85% were discarded. This filtering ended up with 90K segments, out of the 1.7M original segments (approximately 5.3% of acceptance). The resulting segments can be searched using the standard search mechanism of Lucene, which takes into consideration criteria such as rarity of the terms and length of the document [11]. In case no results are found for a word sequence query, the system relaxes the search using each word individually, and then retrieves the most relevant document.

Another difference from the previous sets of exercises has to do with the distractor selection. For each target word of the segment, a distractor is generated. This is the phonetically closest word, chosen from a list of candidates according to a distance metric. The candidate distractors are words from the Portuguese Academic Word List [12]. P-AWL is a set of words specifically developed for REAP.PT. This list is composed of words that the student should learn during the learning process. The current version contains the inflections of about 2K different lemmas, totaling 33.3K words. The *leia* grapheme-to-phone tool [13] was used to obtain the phonetic representation of both the distractor candidates and the target words. Then, the Levenshtein distance was used to determine the closest distractor for each target word. To better represent the distance of two words, a different weight was assigned to each substitution involving a pair of phones [14]. These weights take into account features such as voiced/unvoiced, manner and place of articulation, etc.

As shown in Table 1, set 4 is the only one that requires memorization of the utterance, since the word list is not shown during the playback. On the contrary, in set 5, the user can select words while watching the clip. Finally, the last set introduces two new features. The first one consists in searching a target word in a subset of the corpus, which is comprised of the most recent news, covering the first three months of 2011. The second feature, called *karaoke*, is available at the correction screen, and allows the user to watch the video with the corresponding transcription, while the words are being highlighted as they are spoken.

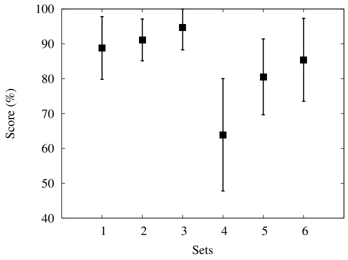

Figure 3: *Average scores obtained by the thirteen participants, for the six sets.*

4. Results

Thirteen non-native Portuguese speakers, from various nationalities and various L1, engaged in the exercises. The average contact with Portuguese was 3.42 years, with a standard deviation of 2.22 years. Ten users have less than 5 years of contact with EP, and three users more than 5 years. The next subsection will analyze the answers to the exercises, concerning the scores, the number of playbacks, etc. Subsection 4.2 will focus on the preferences questionnaire.

4.1. Exercise answers analysis

Figure 3 shows the average scores with standard deviations obtained by the thirteen participants, for each of the six sets. Scores from sets 1–3, and sets 4–6 cannot be compared since the scoring is different for each group of sets. Nevertheless, a trend common to both groups can be found in the positive slope of the scores, showing that the users benefit from exposure to the first examples, most likely by effectively getting used to the two different interfaces and their respective functionalities. Also, set 4 appeared to be the most difficult set, with an average score of 63.9%. This can be explained by the fact that memorization makes this task much more difficult. The first two sets of each interface (sets 1 and 4) have larger standard deviation values , indicating that the users behave more similarly when already accustomed to the interfaces.

The average number of playbacks was 3.4 times per exercise. The first three sets present a higher number of playbacks than the last three, with average values of 4.1 and 2.6, respectively. This can be explained by the fact that the task of sets 1 to 3 involved all the words in the sentences, whereas in the other sets only a subset of words had to be identified. Concerning the slow down feature, results show that it was used 1.1 times on average. However, 60% of the users, who mentioned the speech rate as a difficulty, did not even try this option when it was available. This shows that the way of using this functionality has not been made sufficiently clear. For the exercises where memorization was needed (set 4), users played the videos 2.4 times in average, versus 2.6 times for the other exercises. Although not statistically significant, results showed a tendency for playing back the videos more often when the words that are to be chosen are shown at the same time as the video.

Important conclusions can also be drawn from the answers to the two questions that were asked in each exercise (Section 3). Regarding difficulties, the most signaled by the users was the speech rate, which was selected in 26.5% of the exercises, followed by the sentence length factor (16.3%) . Unknown words were considered as a difficulty only in 8.1% of the exercises. The question that aimed to test the students sensibility to errors also provided interesting results. Users mentioned the existence of errors in 8.2% of the exercises in the sets 1 to 3 – the error-free, manually transcribed sentences. For the other sets, with automatic transcriptions, 64.7% of the sentences with errors were correctly flaged by the users. Most unidentified errors consisted only of minor ones, frequently a deletion or an insertion of a function word. Severe errors, such as misrecognised content words and grammatically incorrect utterances, occurred in 16.3% of the exercises, of which 77.3% were correctly identified as errors. In sum, users seem sensitive to the presence of errors that were due to the automatic processing of the material.

4.2. Questionnaire analysis

Twelve questions were submitted to the users, at the end of the exercises. A five-point Likert scale (1: Completely agree – 3: neither agree, nor disagree – 5: Completely disagree) was used.

Students agreed that they prefer to put all the words in the correct order, rather than to tick some of them in an unsorted list (with a 3.9 value on the scale).

Video was consensually judged as a positive, pleasant and useful feature, with a 3.6 value. Users also found that watching the anchor speak was more helpful than watching an outdoor scene corresponding to what is being said. This could be explained by the fact that the anchor speech is prepared, hence easier to understand than spontaneous speech. Additionally, seeing the mouth and lips of the anchor may also help listening comprehension. The synchronized highlighting of the transcription, *karaoke*, was judged as a helpful feedback feature too. Also consensual, with a 3.9 value, was the preference accorded to being able to solve the exercise at the same time the utterance is being played. Some users even pointed out that memorizing the utterances distracted them from the main goal.

Users agreed that real and recent BN content is an additional motivation. The scores provided after each exercise were also considered to be a positive challenging feature, with a value of 3.7. On the contrary, with an average value of 3.0 and a standard deviation of 1.4, the search feature did not achieved consensus. Users commented that providing suggestions would be an advantage since coming up with words to search is difficult.

Although the slow down feature has not been extensively used during the exercises, this functionality was considered as a benefit, with a 3.5 value on the Likert scale. Users with less than 5 years of contact with EP, found this feature more important than the more experienced users. Finally, as expected, users with longer contact with EP did not consider the sentences as being as difficult as did those with less contact time.

5. Conclusions

By analyzing the results of both the exercises and the questionnaire, it is possible to conclude that, in a future listening comprehension game, adding videos in all exercises, featuring recent content and preferably using anchor speech would constitute positive features. The search engine should include a suggestion mechanism, and the exercises be solved at the same time the clips are being played back. It was also shown that users are sensitive to recognition errors. This suggests the inclusion of an extra task, consisting in asking the users to correct the automatic transcriptions, in order to get more points. This score-giving strategy is an important feedback technique, just as the *karaoke* feature, and may improve students' motivation.

The exercises' level of difficulty, as perceived by the user, varies according to the amount of contact time s/he had with the language being learnt, and so do the students' preferences too. Hence, building a cohesive educational game implies defining a student model. Features such as sentence length, word level, and speech rate, should be weighed in when retrieving exercises for a particular student. At a beginner level, focus should be put in the listening comprehension of a subset of words from an utterance that can be slowed down, whereas at a more advanced level exercises involving the complete set of words of a target sentence could be considered.

6. Acknowledgements

This work was partially supported by FCT (INESC-ID multiannual funding) through the PIDDAC Program funds and by FCT project CMU-PT/HuMach/0053/2008.

7. References

[1] L. Marujo, J. Lopes, N. Mamede, I. Trancoso, J. Pino, M. Eskenazi, J. Baptista, and C. Viana, "Porting REAP to European Portuguese," in *Proc. SLaTE*, Birmingham, 2009, pp. 69–72.

[2] J. Lopes, I. Trancoso, R. Correia, T. Pellegrini, H. Meinedo, N. Mamede, and M. Eskenazi, "Multimedia Learning Materials," in *Proc. IEEE Workshop on Spoken Language Technology SLT*, Berkeley, 2010, pp. 109–114.

[3] J. Richards, "Listening comprehension: Approach, design, procedure," *TESOL quarterly*, vol. 17, no. 2, pp. 219–240, 1983.

[4] T. Secules, C. Herron, and M. Tomasello, "The effect of video context on foreign language learning," *Modern Language Journal*, vol. 76, no. 4, pp. 480–490, 1992.

[5] P. Brett, "Multimedia for listening comprehension: The design of a multimedia-based resource for developing listening skills," *System*, vol. 23, no. 1, pp. 77–85, 1995.

[6] B. Sørensen and B. Meyer, "Serious Games in language learning and teaching – a theoretical perspective," in *Proc. Digital Games Research Association Conference*, 2007, pp. 559–566.

[7] B. Yoshimoto, I. McGraw, and S. Seneff, "Rainbow Rummy: a Web-based game for vocabulary acquisition using computer-directed speech," in *Proc. SLaTE*, Birmingham, 2009, pp. 5–8.

[8] L. Grace and M. Castaneda, "Polyglot Cubed: a Multidisciplinary Listening Comprehension and Recognition Tool," in *Proc. SITE*, Chesapeake, 2011, pp. 3219–3223.

[9] R. Amaral, H. Meinedo, D. Caseiro, I. Trancoso, and J. Neto, "A Prototype System for Selective Dissemination of Broadcast News in European Portuguese," *EURASIP Journal on Advances in Signal Processing*, p. 11, 2007.

[10] H. Meinedo, D. Caseiro, J. Neto, and I. Trancoso, "AUDIMUS.media: A Broadcast News Speech Recognition System for the European Portuguese Language," in *Proc. PROPOR*, Faro, 2003, pp. 9–17.

[11] S. Robertson and H. Zaragoza, "The probabilistic relevance framework: BM25 and beyond," *Information Retrieval*, vol. 3, no. 4, pp. 333–389, 2009.

[12] J. Baptista, N. Costa, J. Guerra, M. Zampieri, M. de Lurdes Cabral, and N. Mamede, "P-AWL: Academic Word List for Portuguese," in *Proc. PROPOR*, Porto Alegre, 2010, pp. 120–123.

[13] L. Oliveira, C. Viana, and I. Trancoso, "DIXI - Portuguese Text-to-Speech System," in *Proc. Eurospeech*, Genoa, 1991.

[14] S. Paulo and L. C. Oliveira, "Multilevel annotation of speech signals using weighted finite state transducers," in *Proc. IEEE Workshop on Speech Synthesis*, Santa Monica, 2002, pp. 111–114.