Automatic generation of listening comprehension learning material in European Portuguese

Thomas Pellegrini[1], *Rui Correia*[124], *Isabel Trancoso*[12],
Jorge Baptista[3], *Nuno Mamede*[12]

[1]INESC-ID Lisboa, Portugal
[2]IST, Lisboa, Portugal
[3]Universidade do Algarve, Portugal
[4]Language Technologies Institute, Carnegie Mellon University
{Thomas.Pellegrini,Rui.Correia,Isabel.Trancoso,Nuno.Mamede}@inesc-id.pt, jbaptis@ualg.pt

Abstract

The goal of this work is the automatic selection of materials for a listening comprehension game. We would like to select automatically transcribed sentences from recent broadcast news corpora, in order to gather material for the games with little human effort. The recognized words are used as the ground solution of the exercises, thus sentences with misrecognitions need to be filtered out. Our experiments confirmed the feasibility of the filter chain that automatically selects sentences, although harder confidence thresholds may be needed. Together with the correct words, wrong candidates, namely distractors, are also needed to build the exercises. Two techniques of distractor generation are presented, either based on the confusion networks produced by the recognizer, or on phonetic distances. The experiments confirmed the complementarity of both approaches.

Index Terms: CALL, Listening Comprehension, European Portuguese, ASR, distractors

1. Introduction

Listening comprehension is an important skill to master, while acquiring a new language. Portuguese has one of the richest phonology among the Latin languages. It has a large set of vowels with oral and nasal vowels, oral and nasal diphthongs and double diphthongs, an example of the latest being the very common first-name João, [ʒoɐ̃w̃]. The European Portuguese (EP) variety distinguishes itself from other varieties, in particular from Brazilian Portuguese, by strong vowel reduction in unstressed syllables. Unstressed vowels are either centralized or simply omitted [1, 2]. This characteristic contributes to make spoken EP particularly difficult to understand for non-native speakers. This fact, combined with the scarcity of learning materials for EP, hinders the process of providing listening comprehension exercises to help non-native learners developing this specific skill. For that reason, building Computer-Assisted Language Learning (CALL) tools to enhance listening comprehension is one of our current priorities.

Our first efforts in this direction, were to introduce multimedia materials in the Portuguese version of REAP [3, 4]. REAP.PT began as a tool oriented for vocabulary acquisition. An important research topic in such systems, is the generation of *fill-in-the-blanks* questions – often referred to as *cloze* questions [5]. This involves not only the generation of the questions themselves, but also of distractors. For EP, we explored a phonetic strategy [6], which generated misspelled words, phonetically close to a specific target word, endowing the exercise with an additional spelling component. Another innovation of REAP.PT was the integration of a listening comprehension module, which allows the students to watch the last day's news, while simultaneously reading the automatically transcribed subtitles [7]. This research direction led us to explore games to further stimulate listening comprehension skills.

Serious games have recently gained strong interest in the CALL community to support L2 acquisition. These games have an objective of educational design and beyond entertainment [8]. They range from puzzles and minigames similar to the well-known Hangman games, to more sophisticated video games, such as Mingoville[1], and Polyglot [2].

The goal of this work is the automatic generation of learning material for listening comprehension games in EP, in order to avoid or at least to ease the manual supervision for this task. Game materials consist of audio sentences along with their transcriptions. The students listen to the utterances, and try to identify all or part of the spoken words. To avoid inhibition due to word spelling, word candidates, both correct words and distractors, are presented to the students as puzzle pieces that must be correctly ordered to form the original sentence. Figure 1 gives an example of one of our games.

In the literature, many papers focus on the automatic generation of exercises and distractors oriented to assess vocabulary and grammar. These studies rely on natural language processing techniques, using syntactic and semantic features to accomplish their goals. In this paper, we report on the automatic selection of speech utterances, by developing a filter chain, which discards unsuitable sentences, such as those with misrecognitions. Distractors were generated with two techniques: an ASR-based approach and a phonetic-based one, following up on our experience in REAP.PT.

The paper is organized as follows. Next section describes the filter chain that selects speech sentences from broadcast news (BN) shows. Section 3 focusses on distractor generation, describing a technique based on the ASR confusion networks, as well as a phonetic-based technique. Finally, evaluations of the quality of the filtering and of the generated distractors, are proposed in Section 4.

[1]www.mingoville.com (visited in May 2011)
[2]www.polyglotgame.com (visited in May 2011)

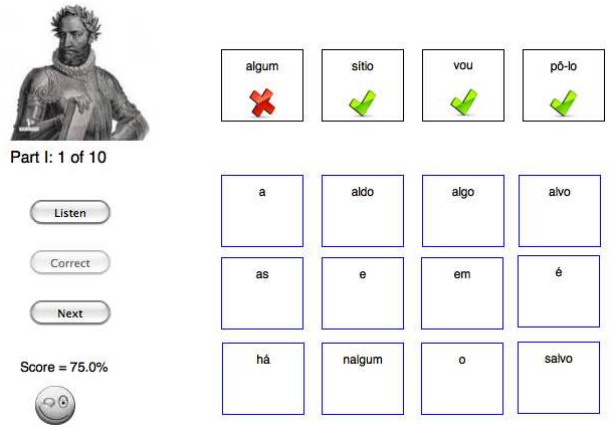

Figure 1: *Game interface for the utterance "nalgum sítio vou pô-lo"*

2. Sentence constraints

Broadcast news can be a motivating material in the context of L2 acquisition for adults, because of the variety of topics covered, the ever updating content, and the possibility to understand better the culture of the country. Daily Portuguese broadcast news are automatically transcribed by our ASR engine, named "Audimus", an hybrid Hidden Markov Model - Multilayer Perceptron decoder [9]. This material provides sentences that could be used in listening comprehension games. However, automatically transcribed sentences may contain misrecognized words and may be incomplete. The audio quality also needs to be checked. For example, an outdoor report may be noisy and is more difficult to understand, when compared to anchor speech. Therefore, a sequence of five filters, described hereafter, was applied to the automatically transcribed sentences.

2.1. Sentence length - *f1*

In [10], Ur stressed out the necessity of providing small units of speech for listening comprehension assessment, since in real-life, the discourse is usually divided in small chunks of speech. Sentences with a minimum of 4 words and a maximum of 10 words were selected. From our experience, shorter sentences would be too easy, and longer ones too difficult. It can be noticed that in BN shows, mainly comprised of prepared speech, sentences tend to be longer.

2.2. ASR confidence measure - *f2*

Confidence measures (CM) at word-level allow to estimate how reliable is each hypothesized word. Computed within a $[0, 1]$ range, CM are usually estimated by gathering scores and other informative variables, during the ASR decoding [11]. For this study, CM at sentence-level were used as a sentence filter. We computed the average CM of the words that compose the sentence. Only sentences with CM larger than 0.9 were selected.

2.3. Syntactic completeness - *f3*

Utterances with at least one verb, and one common noun or one adverb, were selected. This filter prefers sentences that contains content words, and not only functional words.

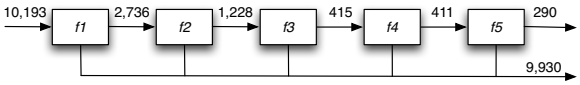

Figure 2: *Filter chain application*

2.4. Signal-to-noise ratio - *f4*

The signal-to-noise ratio (SNR) is an important feature in determining the quality of audio data. ASR performance is strongly influenced by the SNR. We used a short-term analysis of the speech signal with 20ms frames[3].

Utterances with a SNR of less than 10dB were rejected. It allowed to reject speech over music utterances, that mainly occur in the beginning of the BN shows (headlines).

This filter is expected to reject a small number of utterances, since for small SNR values, i.e. with high energy noise, the ASR performance decreases rapidly, hence confidence measures for hypothesized words are expected to be low. The *f2* CM constraint should already filter out most of these cases.

2.5. Neutral declaratives - *f5*

End of sentences are determined by a statistical module that recovers full-stops and commas [12]. A recent extension of this module allows the recovery of question marks, but since interrogatives are very rare in broadcast news, we decided to segment utterances by using only full stops.

In order to reject speech continuations, and give preference to neutral declaratives, an additional constraint on pitch was used. In European Portuguese, neutral declaratives usually show decreasing or stable pitch pattern endings [13]. Sentences were rejected when the pitch slope of the last voiced segment of the utterances was positive, indicating a potential interrogative sentence. The fundamental pitch was extracted by using the Snack software[4].

2.6. Filter chain application

To generate exercises automatically, the daily Portuguese national BN shows from February 2011 were automatically transcribed, totaling a total of more than 10k candidate sentences.

The application of the filter chain to the set of 10k sentences resulted in very drastic reductions. The length filter, *f1*, reduced by 73% the number of sentences. This can be explained, as already mentioned, by the typically large length of BN sentences (17 words per sentence, on average). Filter *f2*, responsible to filter out sentences with recognition errors. The typical Word Error Rate of our in-house ASR engine is approximately 18% on BN data [14]. Thus, in average, a 10-word sentence is expected to contain 2 misrecognitions. This fact explains the high percentage of the discarded sentences by this filter (55%).

Further reductions were obtained by applying *f3* and *f5*, 66% and 29% respectively. As expected *f4*, did not introduce significant reductions. Figure 2 shows the execution of the filter chain and the filtered sentences at each step. In sum, only 290 out of 10k sentences (3%), passed all filters.

[3]http://www.isip.piconepress.com/projects/speech
[4]http://www.speech.kth.se/snack

3. Distractor generation

Two techniques for generating distractors are described in this section: distractors gathered during the ASR decoding (ASR-based), and distractors selected according to phonetic distances (phonetic-based).

3.1. ASR-based distractors

During the decoding process, various word hypotheses are competing. Intrinsically, the words in competition are phonetically similar, and as such, it may be difficult for students to distinguish between them. One innovative aspect of this study takes advantage of this fact, by using these words as distractors for listening comprehension exercises.

The best hypothesis, which corresponds to the lowest score path inside a word lattice or a confusion network, is the ASR transcription, and therefore is the solution of the exercises. All the other competing words generated during the decoding, were used as distractors. Confusion networks, stored during the decoding, permit to easily associate each best hypothesized word with its competing words.

The number of competing words depends on the beam search and on the maximal number of arcs parameters used in the decoding process. It also depends on the overall difficulty to transcribe the utterance. For well-formed utterances (no disfluencies, clean audio, prepared speech, etc.), the decoder does not hesitate between various hypotheses. In this case, no or very few distractors can be derived from this method.

3.2. Phonetic-based distractors

This method uses the final output of the ASR, generating a fixed number of two distractors for each recognized word. As the name states, these distractors are selected exploring the phonetic representations of each recognized word.

Candidate distractors are words from the ASR vocabulary (100k words). This vocabulary is updated everyday, although maintaining its size, by dropping very infrequent words and adding new ones as they appear more often in the news. For each recognized word, only the two closest candidates, in terms of a phonetic distance, are selected as distractors.

To compute the phonetic distance between two words, the Levenshtein distance was used between the phonetic representations of each pair of words (correct word and candidate distractor). These representations were obtained using the *leia* grapheme-to-phone tool [15].

However, if one simply used the distance algorithm directly over the phonetic representations of the two words to compare them, we would end up with an approach similar to orthography comparison. Thus, a different weight was assigned to each substitution between a pair of phones. These weights were based on Paulo and Oliveira's work [16], which took into consideration features such as voiced/unvoiced, manner and place of articulation, etc. The weights for deletion and insertion are 10 and 11, respectively. Table 1 shows a fragment of the weights that were used for the substitution operation.

In sum, the phonetic-based distractors for a given sentence are comprised by the two closest candidate distractors of each word of the recognized sentence.

As an example, for the word "começo" ([kumesu]) – meaning *beginning* – the generated distractors were "comesse" ([kum"esə]) – a form of the verb *to eat* – and "conheço" ([kuɟ"esu]) – a form of the verb *to know* - as the first and second closest options, respectively.

Table 1: *Fragment of the weights for substitution operation*

Base Phone	Target Phone	Distance
[i]	[i]	0
[i]	[e]	4
[i]	[u]	6
[i]	[ɐ]	8
[i]	[ɔ]	10

4. Evaluation

To evaluate both the quality of the filtered sentences and the quality of the distractors, a set of 80 sentences was used: 40 sentences from manual transcriptions and 40 sentences randomly chosen among the 290 automatic filtered sentences. The 40 manually transcribed sentences were manually selected, according to the same criteria used to define the filters designed for the ASR material. The filter chain was applied afterwards to this manual set (for the CM filter *f2*, a CM equal to 1 for each word was assigned). None of the sentences were discarded by the filters, showing a good adequacy between the subjective criteria and the filter chain.

Two Portuguese native speakers, with backgrounds of spoken language processing (annotator 1), and linguistics (annotator 2), answered a survey about the 80 sentences. The objectives of the survey were the following: first to see if automatically transcribed sentences would be discriminated from manual transcriptions, allowing to validate or not the use of ASR material, second to identify the problems for which sentences could be rejected, and finally to see if the two automatic methods to generate distractors could be used to suggest distractors.

For each sentence, the transcription, with possible errors for the ASR sentences, was shown, and the annotators had to listen to the corresponding speech utterance. They were asked to rate the "general quality" of each sentence on a five-point Likert scale (1=*very bad*, 3=*OK*, 5=*very good*). If the answer was not *very good*, problems could be identified among a list of five choices: *ASR errors*, *noise*, *syntactic*, *semantic*, *other*. Comments could also be added. A *syntactic* problem corresponds to a grammatical error or uncorrectness. A *semantic* problem corresponds to a misrecognition that makes the sentence nonsense. Finally, for each sentence, annotators were asked to pick up distractors among a list of ASR-based and phonetic-based distractors, with no information about the method used to generate them.

4.1. Filter evaluation

Average scores for the automatic set were smaller than for the manual set: respectively 4.32 and 4.97 for annotator 1, 4.41 and 4.70 for annotator 2. Both annotators rated 60% of the sentences with the same score. Almost all the sentences for which the annotators agreed, were ranked as *very good*. It is interesting to notice that the two annotators evaluated the quality differently. In general, annotator 1 was more exigent with the ASR quality than annotator 2, assigning respectively 3-valued (*OK*) and 4-valued (*good*) scores to sentences with small ASR errors. For almost all the sentences marked by annotator 1 with a quality lower than 5, an *ASR error* problem was pointed out. Common ASR errors are due to co-articulation effects, such as "justificou o Estado," which was recognized as "justifica-o Estado," which is syntactically incorrect. A rule-based syntactic filter may detect this type of errors involving clitics. In fact, with the

exception of one ASR error on a content word, all the ASR errors correspond to deletions, insertions, or substitutions of small functional words. Annotator 2 ranked these sentences with a 4-valued or 5-valued quality score. Both annotators ranked as *bad* the quality of the sentence with an ASR error on a content word.

Another difference concerns the perception of noise. For annotator 2, the presence of noise in both the automatic and the manual sets was judged as a limitation, whereas for annotator 1, noisy examples are interesting to show to the learners, as real-life speech examples, if the SNR is large enough. This evaluator did not mark any sentence with the *noise* problem.

These results show small but significative differences in subjective quality rating between the automatic and the manual sets. Most common problems concern ASR errors, involving small functional words. The CM threshold used in filter *f2* could be larger, and completed by other filters, such as syntactic rule-based filters.

4.2. Distractor evaluation

In total, more than 1k distractors were generated for the 80 sentences. ASR-based and phonetic-based distractors totalized 377 and 697 respectively. As explained in section 3, ASR-based distractors are all the words competing during the decoding. For the phonetic-based distractors, two distractors per reference word were used.

Annotator 1 selected 45% of the distractors in total, both ASR- and phonetic-based words, much fewer distractors than Annotator 2, who selected 60% of them. Both annotators agreed on 57.0% of the distractors.

Annotator 1 privileged ASR-based distractors, by choosing 46.1% of them versus 44.9% of phonetic-based distractors. Annotator 2 chose a slightly larger rate of phonetic-based distractors, 60.7% versus 59.0% of ASR-based distractors. These percentages do not sum up to 100%, since each corresponds to the ratio of the number of selected distractors of one type (ASR-based or phonetic-based distractors), divided by the total number of proposed distractors of the same type.

Both distractors types have been chosen in similar proportions by both annotators, showing a complementarity in the two generation methods. It is also interesting to notice that when both ASR- and phonetic-based approaches generated the same distractor (which only occurred for 3.5% of the distractors), it was selected 61% and 92% of the time, by annotator 1 and annotator 2, respectively.

ASR distractors present the advantage to cover co-articulation effects, i.e. they may correspond to multi-word common confusions, that make them very interesting. On the other hand, when the ASR decoder is pretty confident in an hypothesis, no ASR-based distractors exist. In this case, the sentence may be filtered out, or only phonetic-based distractors may be used. It will be interesting to analyse whether it is worth generating phonetic-based distractors for function words or proper nouns.

5. Conclusions

The goal of this work was the automatic selection of materials for an listening comprehension game. We would like to select sentences from recent BN corpora which are automatically transcribed, and to present as puzzle pieces for the student to pick the one-best words that are hypothesized by the ASR system together with distractors. Our experiments confirmed the feasibility of the filter chain that automatically selects sentences, although harder confidence thresholds may be needed. The experiments also confirmed that automatically generated distractors are feasible but a detailed analysis of the distractors rejected by both annotators is needed. The complementarity of the two distractor generation methods is worth emphasizing.

Several parameters in the game may be tuned to the level of the student: the sentence length, the CM and SNR thresholds, the minimum lexical frequency of the candidate words, etc. Slowing down the sentence playback specially for lower level students is also a topic for future work.

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. The authors would also like to thank Maxine Eskenazi for many helpful suggestions.

7. References

[1] M. Cruz-Ferreira, "European Portuguese," *Journal of International Phonetic Association*, vol. 25:02, pp. 90–94, 2009.

[2] J. Rouas, I. Trancoso, C. Viana, and M. Abreu, "Language and variety verification on broadcast news for Portuguese," *Speech Communication*, vol. 50, no. 11-12, pp. 965–979, 2008.

[3] 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*, Pittsburgh, 2006, pp. 829–832.

[4] 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.

[5] J. Pino and M. Eskenazi, "Semi-Automatic Generation of Cloze Question Distractors Effect of Students' L1," in *Proc. SLaTE*, 2009.

[6] R. Correia, J. Baptista, N. Mamede, I. Trancoso, and M. Eskenazi, "Automatic Generation of Cloze Question Distractors," in *Proc. SLaTE*, Tokyo, 2010.

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

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

[9] H. Meinedo and J. Neto, "Audio Segmentation, Classification and Clustering in a Broadcast News Task," in *Proc. ICASSP*, vol. II, Hong Kong, 2003, pp. 5–8.

[10] P. Ur, *Teaching listening comprehension*. Cambridge University Press, 1998.

[11] T. Pellegrini and I. Trancoso, "Improving ASR error detection with non-decoder based features," in *Proc. Interspeech*, Makuhari, 2010, pp. 1950–1953.

[12] F. Batista, H. Moniz, I. Trancoso, H. Meinedo, I. Mata da Silva, and N. Mamede, "Extending the punctuation module for European Portuguese," in *Proc. Interspeech*, Makuhari, 2010, pp. 1509–1512.

[13] S. Frota, *Prosody and Focus in European Portuguese. Phonological Phrasing and Intonation*. New York: Garland Publishing, 2000.

[14] H. Meinedo, A. Abad, T. Pellegrini, I. Trancoso, and J. Neto, "The L2F Broadcast News Speech Recognition System," in *Proc. Fala 2010*, Vigo, 2010.

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

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