# Automatic Generation of *Cloze* Question Distractors

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

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

## Abstract

This paper presents a technique to generate distractors for *cloze* questions in the context of a Computer-Assisted Language Learning tutoring system. The document will focus on an evaluation process used to measure the quality of the distractors that were automatically generated. The main goal of the present study is to be able to include this feature in the tutoring system.
**Index Terms**: automatic distractors generation; *cloze* questions; Computer Assisted Language Learning; Portuguese.

## 1. Introduction

REAP.PT [1] (REAder-specific Practice PorTuguese) is the Portuguese version of the REAP system [2] developed at Carnegie Mellon University. As the name states, REAP.PT is based on the importance of reading activities as a way to become proficient in a new language. From the standpoint of the student, the learning method can be summarized in two main phases: text reading and question answering.

The paper focus on the second phase. It explores a way of automatically generating questions, more specifically, generating distractors for a specific type of questions: *cloze* questions. In a *cloze* question, the testee is asked to find the word, among a set of answer choices, that better fits in a specific stem (a sentence with a blank space).

The use as distractors in *cloze* questions, as opposed to using open questions, meets the requirement of having a system with a certain degree of automation in the learning process, not requiring the manual grading of answers by the teacher. This would have too much impact in the dynamics of the system, slowing down the acquisition of vocabulary by the students. On the other hand, having a proper set of distractors associated to each question may also work as a guide, driving the student's attention into a specific and controlled set of words.

Generating this type of question, in the REAP.PT context, implied two main resources: in the first place, a list of words that students should learn (i.e., the words that are going to be tested in the questions), and in the second place, a set of *stems* that will be used to test the vocabulary acquisition. The first resource is the Portuguese Academic Word List [3] (from now on, P-AWL) – a vocabulary specially designed for language learning. P-AWL is "a careful selection of common words that may constitute a valid tool for assessment of language proficiency at university level, irrespective of scientific or technical domain. One can view P-AWL as a landmark, useful to measure the students' progress on his/her learning process and language proficiency." The current version is composed by 2,019 different *lemmas*, together with their most common inflections, each one tagged with the correspondent part-of-speech (POS), totalizing 33,284 words. Verbal forms represent 88% of the extended P-AWL, given that most verbs include 68 inflected forms.

The second resource is a set of 6,000 sentences that were selected and adapted by linguists to serve as the basis for a question generation module of the REAP-PT system. These sentences were selected from text and web corpora, according to a predefined set of criteria: **(i)** only full sentences and not fragmentary text; **(ii)** no titles, captions, and other paratextual elements; **(iii)** no definitions and other lexicographic context; **(iv)** the target word should not be at the beginning nor at the end of the sentence; **(v)** sentences should be short but not too short, between 100 to 200 characters; **(vi)** in the case of ambiguous words, different meanings are represented by independent sentence sets; **(vii)** sentences should correspond to a "natural" or "characteristic" distribution of the target word; **(viii)** sentences should constitute a non ambiguous environment for the correct identification of an ambiguous word; **(ix)** sentences should provide a balanced set of each target word, and its most current inflected forms.

This document will present the methods used to generate distractors, compare them and derive conclusions concerning their eventual integration in REAP.PT.

## 2. Related work

Graesser and Wisher [4] present directives for multiple-choice questions. Their work proves that the ideal number of distractors is three plus the correct answer.

Goodrich [5] presents a way to determine the efficiency of distractors. To measure it, two concepts are involved: *potency* and *discrimination*. *Potency* is the percentage of students that make a specific choice. Here, there is a trade-off involved between having nobody choosing a particular distractor, indicating that it is not "giving the question a factor of difficulty", and being frequently selected, indicating that it may be a "correct answer to a badly posed question". *Discrimination* has to do with the ability to differentiate students of different levels of proficiency. Goodrich also distinguishes some distractors categories. Some of them, like the use of antonyms and false synonyms (words that have close or similar meaning but cannot fit in the context of the stem) are fragile regarding automatic generation, since they are very context dependent, and can easily lead to the generation of correct choices instead of distractors. Some of the categories discussed in Goodrich's work were used for

This work was partly supported by FCT (INESC-ID multiannual funding) through the PIDDAC Program funds and the REAP.PT project.

the present evaluation, namely random distractors, graphemic variation and morphological variation.

Pino and Eskenazi [6] focused on this same subject for the English version of REAP. In their work, each one of the 33 *cloze* questions was assigned to a set of distractors, each one belonging to a different category (morphological, orthographic, phonetical, and combinations of orthographic-morphological and phonetical-morphological). This study was aimed at non-native speakers, relating the origin of the student with the category that proved to work better as a distractor. In fact, the native language of the student proved to influence the distractors choice.

## 3. A preliminary experiment

A preliminary experiment was carried out among 4 test subjects, using a set of 100 randomly selected sentences (from the already discussed corpus of 6,000 sentences), distinct from the set used for the main experiment. They were asked to complete each sentence, not being provided any answer choices. With this experiment, one expected a low percentage of correct answers (answering exactly the target word that a specific stem aims to test), and were interested in finding the most common reasons for "incorrect" answers.

The subjects did not participate in the test in exactly the same conditions: **(i)** subject A had no previous knowledge of the sentences; **(ii)** subject B does not consider her/himself as native-speaker proper, and had only partial knowledge of the corpus of sentences, since s/he has been involved in their selection; **(iii)** subject C and D are Portuguese native speakers and they also had only partial knowledge of the corpus of sentences.

Since subjects B, C and D were involved in the selection of sentences, a random set was retrieved from the sentences selected by each one, so that only 25% of the sentences were in fact previously known to each person. For each response, there were three possible outcomes: unable to find a coherent word to complete the stem, able to find exactly the expected word and able to find a word but not the expected target word. Results are shown in Table 1:

| | A | B | C | D |
|---|---|---|---|---|
| No Response | 4 | 23 | 23 | 24 |
| Correct | 2 | 14 | 27 | 20 |
| Incorrect | 94 | 63 | 50 | 56 |

Table 1: *Distribution of the answers.*

While subjects B, C and D, with previous, even if partial, knowledge of the corpus of sentences were able to provide correct answers for 14, 27 and 20 sentences, respectively; subject A only got 2 answers right. This seem to confirm that previous knowledge of the sentences, even if diluted among the large number of each subject's selected sentences and the time gap between the selection and the testing (over a few months), may influence their response.

The attitude towards the test has also been different among the participants. Subject A only left unanswered 4 questions, while B, C and D were much less assertive or were in fact unable to find adequate answers for 23 or 24 questions.

The analysis of the "incorrect" answers is illustrative of the cognitive mechanisms involved in the test and may shed some light on both the quality of the sentences and the task difficulty. It should be kept in mind that having chosen a word that is not the target word does not mean that it is inadequate.

This study revealed some problems associated with the use of open *cloze* questions (when no set of answer choices is provided). While it would be impossible to go through all cases here we will highlight some remarks: **(i)** "incorrect" answers may arise from the choice of synonyms or antonyms (notice the case where the target word is semantically neutral (for example "vary"), and incorrect answers correspond to the positive ("increase") and negative ("reduce") polarity); **(ii)** hyponymy and hyperonymy relations often compete with synonyms/antonyms as "incorrect" answers; **(iii)** in some cases, a specific full verb is replaced by a support (or "light") verb and corresponding nominalization, practically devoid of meaning (for example, "give" instead of "confer"); **(iv)** less frequent adverbial quantifiers (like "marginally") are replaced for equivalent adverbs but with a broader selection ("very" and "little"); **(v)** collocation patterns arise in the choice of the answers that do not match the target word.

The use of multiple-choice questions with automatic distractor's generation is able to solve some of the aforementioned problems, such as the use of synonyms, antonyms, hyponyms and hyperonyms as answers (using, for instance, lexical resources to eliminate these choices).

## 4. Experimental Setup

For this experiment, 20 stems were randomly selected and, for each, six sets of distractors were generated (each element of a given set was generated with the same generation strategy as the remaining ones), thus yielding 120 sets of distractors. This setup allows one to isolate each generation method thus being capable of draw conclusions over each method, separately. Each test subject was asked to answer 10 *cloze* questions. The questions were randomly chosen, maintaining a balance over the pairs stem-distractors that were already answered, and avoiding the use of repeated stems in each test.

The test subjects, also divided by native and non-native speakers, were asked to select all the words that could fit in the stem from the set of words provided. We had 247 participants in our test (212 native and 35 non-native speakers).

The present study focus on 4 main generation methods of distractors: manual, random, graphemic, and phonetic distractors. For the random and graphemic methods we used lexical resources in order to filter out synonyms, hyponyms and hyperonyms of the target word, which, if included, might also fit the stem as correct answers and fail to function as distractors. We will now focus on each method in particular.

### 4.1. Manual Distractors

For a set of 20 randomly selected stems, a set of distractors was manually produced by the team members. These distractors were selected based on the following criteria: **(i)** quasi-synonymous ou quasi-antonymous words, which do not correspond exactly to the negative/positive overall sense of the target word in the sentence; **(ii)** similar spelling or similar sounding words; **(iii)** false-friends, taking as competing languages the pair English/Portuguese; **(iv)** (pseudo-)prefix and suffix variation.

For example, for the target word *condução*, in the sense of 'driving a vehicle', the manually selected distractors were: **(i)** *direcção*: noun derived from the verb *dirigir* (drive) that is a perfect synonym of *conduzir* from which the target word is derived (notice, in this case, that even if the two verbs are synonyms, their derived nouns are not, and cannot fit in the stem);

**(ii)** *condição*: phonetic/graphemic similarity; and **(iii)** *redução*: pseudo-prefix re-/con- variation, but otherwise semantically unrelated words.

### 4.2. Random Distractors

This method was developed to set a baseline on automatic distractor generation. Despite the name, this method is not completely random. Coniam [7] recommends the use of distractors within the same POS and frequency rate of the word that is being tested.

The distractors were selected among the P-AWL entries, which were tagged with the corresponding POS (including gender/number/tense/mode/person information). Since distractors are being generated for a system that tracks the evolution of the student over the vocabulary learning process, word level was used instead of frequency rate. Word level results from assigning to each P-AWL entry a predefined value. Ideally, this value would represent the complexity of the word in the Portuguese language, as it is reflected by the school year this word is introduced and the year at which its mastery is deemed to be achieved. Unigrams were used to measure the probability of a word belonging to a certain grade level. A corpus of text books from Porto Editora was used for this purpose. Each textbook was designed for a specific level of secondary school, from 5 to 12. It was thus possible to build unigram language models, and assign a level to each word. The distractors were then chosen using the P-AWL entries with the same POS and level.

### 4.3. Graphemic Distractors

To compute this set of distractors, the P-AWL words with the same POS were used, and the Levenshtein Distance [8] was computed from the target word to each word, using unit costs for the three operations (deletion, insertion and reversal). The words with the lowest distances were selected, among the ones which provided a distance lower than five. The set of distractors thus obtained was completed by recurring to words with a lower distance from a different POS (ending up with gender or number variations for nouns, or tense and person for verbs).

### 4.4. Phonetic Distractors

This method of generation aims at exploring the most common spelling errors for the Portuguese language. To accomplish this task, a table of common mistakes was used. For example, "ss" is frequently confused with "ç" (before "a", "o" and "u") and "c" (before "e" and "i"); "j" can be in some cases confused with "g" (before "e" and "i"); etc. The target word is submitted to this table of modifications and several (misspelled) words are thus obtained.

The *leia* grapheme-to-phone tool [9] was used to provide phonetic transcriptions, so that misspelled words sharing the same transcription as the target word might be selected.

For example, the Portuguese word *começar* (to start), shares the phonetic transcription (/kum@s"ar/, using SAMPA symbols) with the misspelled words *cumeçar*, *comessar*, etc.

### 4.5. Filtering using lexical resources

Lexical resources may improve methods of automatic distractor generation by filtering out correct candidates choices, other than the target word. However, Portuguese still lacks lexical resources with enough coverage and quality for this task's requirements. Trying to overcome these problems, and having in mind targeting the coverage of the maximum number of words from P-AWL, we used two different resources: Papel and MWN.PT. These two resources were used to generate the *Random + Filtering* and the *Graphemic + Filtering* categories of distractors.

#### *4.5.1. Papel*

PAPEL[1] (Porto Editora's Associated Words – Linguateca) is a free lexical resource that focuses on word relations. It is a relatively recent resource, developed between September 2007 and December 2008. Currently, PAPEL is in version 2.0, available since March 2010. PAPEL supports synonym, hyponym and hyperonym relations and has a recall of 80.5% of the P-AWL vocabulary. PAPEL was used to exclude distractors with this type of relationship with the target word. For instance, for the target word *refinação* (refinement), our graphemic method produced the distractor *realização* (implementation) which was later discarded with the aid of PAPEL. Unfortunately, there is no support for antonym relations with this resource.

#### *4.5.2. MWN.PT*

MWN.PT[2] (MultiwordNet of Portuguese) is a lexical resource shaped under the ontological model of *wordnets*. It also focuses on relations between words. It is available since May 2008, being the first publicly available wordnet for Portuguese. MWN.PT is an ontological model representing relations in a hierarchical manner, in which words are grouped in synonym sets, called "synsets", which establish a relationship of synonymy between those words. MWN.PT has 17,200 concepts, "made of over 21,000 word senses/word forms and 16,000 lemmas from both European and American variants of Portuguese". This resource also provides synonymy, hyponymy and hyperonymy relations, covering only nouns (recall of 60% of the nouns in P-AWL). However, MWN.PT catalogues each word in a specific domain. We used this property to exclude words within the same domain (for example, if two words represent an occupation – "lawyer" and "doctor" – they can easily fit in the same stem, if the context does not restrict the choice).

## 5. Results

Results obtained with different methods were compared with the results achieved with manually produced distractors. The latter may be seen as a sort of goal/best result that those methods would try to emulate. Table 2 presents the distribution of correct answers, i.e., answers in which only the expected choice was selected, across the type of distractors, for both native (NS) and non-native (NNS) speakers.

| Method | NS (%) | NNS (%) |
|---|---|---|
| Manual | 87.9 | 70.0 |
| Random | 83.1 | 78.0 |
| Random + Filtering | 84.3 | 73.8 |
| Graphemic | 83.2 | 61.0 |
| Graphemic + Filtering | 87.0 | 86.0 |
| Phonetic | 88.3 | 75.7 |
| Total | 85.6 | 74.2 |

Table 2: *Percentage of correct answers.*

As one would expect, native speakers tend to achieve a

[1] http://www.linguateca.pt/PAPEL/ (last visited in May 2010)
[2] http://mwnpt.di.fc.ul.pt/features.html#main

higher percentage of correct answers than non-native speakers. This result is consistent for all our distractor generation methods.

Variation due to the different methods does not seem to affect the results of native speakers in a significant way. The standard deviation with native speakers is 2.37 while with non-native speakers is 8.37, that is, about three times higher. This seems to indicate that non-native speakers are more prone to produce incorrect answers depending on the distractor generation method than native speakers, who can activate their language knowledge to detect other clues in the stem in order to find the best match.

Concerning native speakers, the fact that the results for the phonetic method are slightly superior (88.3%) can be easily justified since they are aware of the spelling rules for their language. In fact, having sets of distractors composed only by this method makes it easier for native speakers, since they become really focused on the task of finding the misspelled words. As it would be expected, the lowest result was obtained with the random distractors (83.1%), and even so it is only 4.8% (absolute) less than the manual method.

Results for non-native speakers are on average 11% lower than those for native speakers. The highest difference occurs for the manual distractors (17.9%). This difference may be taken as a confirmation of the adequate selection of the manual distractors. The random and phonetic methods show similar results (78.0 and 75.7%, respectively). However, when both are compared with the performance of native speakers, the phonetic method reveals its interest for distractors generation, since it provides significantly less correct answers (less 5.1 and 12.6%, respectively). The graphemic method exhibits the lowest number of correct answers for non-native speakers (22.2% less than the results for native speakers). This result indicates that the confusion introduced by presenting similar (valid) words may be the most important cause for the incorrect answers, and that this method might be particularly suited for screening different levels of progression in vocabulary acquisition.

For native speakers, the filtering versions of the generation methods always produced higher percentages of correct answers (1.2% more than the random method and 3.8% than the graphemic method). This positive, if small, impact in performance may also means that it became easier (even if only a little easier) to rule out incorrect answers.

It is also important to focus on the number of choices testees selected. Non-native speakers always selected only one choice, a fact that can be explained by their concern to answer the question right, stopping when they have found a correct answer (in their notion). For native speakers we registered a very low percentage of selection of more than one word, across the several types of distractors (0.29% for the random, random filtered with resources and phonetics categories, 0.34% for graphemic filtered with resources and 0.57% for the graphemic ones). The manual category, as expected, did not generate any confusion.

Several native testees commented on the level of the stems themselves. Although they were randomly selected among a set of manually revised sentences, care was not taken to ensure that the level of the sentence with a blank space fitted the level of the target word. In fact, a later analysis revealed that 70% of the sentences had a higher level than the corresponding target words (on average, 2.4 higher than the target word level, with a standard deviation of 1.6). This was a problem for non-native testees, who in some cases had to look up unknown words in a dictionary.

# 6. Conclusions and future work

This paper compared several ways of producing distractors. For native testees, the phonetic approach provided the highest percentage of correct answers, closely followed by the manually produced distractors and the graphemic plus filtering approach. The random (with and without filtering) and the graphemic approaches yielded the lowest results.

For non-native speakers, despite their low representativeness, the highest rate of correct answers was obtained with the graphemic plus filtering approach, and the lowest one with the graphemic approach. The phonetic approach and the two approaches with filtering yielded rates of correct answers which are similar to the one achieved by the manually generated distractors, indicating a good adequacy for the integration of these approaches in the REAP.PT system.

The use of different strategies for automatically generating distractors according to the student's level of comprehension of the language seems useful to guide the student throughout the vocabulary learning process of Portuguese. The false-friends distractors, used in the manual generation, are also an interesting topic for future research. Another one is the inclusion of *collocations*, thus being able to increase the difficulty and correctness of distractors, generating the ones that have low collocation values in a specific stem.

Another future topic for research is the generation of the stems themselves, allowing us to have a completely automated process of generating questions for the REAP.PT system, giving only as an input a set of words that are to be tested. The predefined set of criteria used in the selection needs to be complemented with the correspondence between the levels of the stem and target words.

The work on *cloze* question generation would significantly benefit from the integrating of new/larger resources, with much higher recall rates, and more exhaustive semantic information.

# 7. References


[1] Marujo, L., Lopes, J., Mamede, N., Trancoso, I., Pino, J., Eskenazi, M., Baptista, J. and Viana, C., "Porting REAP to European Portuguese.", Proc. SLaTE Workshop on Speech and Language Technology in Education, 2009.

[2] Heilman, M., Collins-Thompson, K., Callan, J. and Eskenazi, M., "Classroom success of an Intelligent Tutoring System for lexical practice and reading comprehension", Proc. Ninth International Conference on Spoken Language Processing, 2006.

[3] Baptista, J. and Costa, N. and Guerra, J. and Zampieri, M., Cabral, M. L. and Mamede, N., "P-AWL: Academic Word List for Portuguese", Proc. PROPOR 2010, LNAI 6001, pp. 120–123, 2010.

[4] Graesser, Arthur C. and Wisher, R. A., "Question Generation as a Learning Multiplier in Distributed Learning Environments", U.S. Army Research Institute for the Behavioral and Social Sciences (Alexandria, Va), A654993, 2001.

[5] Goodrich, H. C., "Distractor efficiency in foreign language testing", TESOL Quarterly, vol. 11, no. 1, pp. 69–78, 1977.

[6] Pino, J. and Eskenazi, M., "Semi-Automatic Generation of Cloze Question Distractors Effect of Students' L1.", Proc. SLaTE Workshop on Speech and Language Technology in Education, 2009.

[7] Coniam, David, "A Preliminary Inquiry into Using Corpus Word Frequency Data in the Automatic Generation of English Language Cloze Tests", CALICO Journal, vol. 14, no. 2, pp.15–33, 1997.

[8] Levenshtein, VI, "Binary codes capable of correcting deletions, insertions, and reversals", Soviet Physics Doklady, no 10, pp. 707–10, 1966.

[9] Oliveira, Luís C., Viana, C. and Trancoso, I., "DIXI – Portuguese Text-to-Speech System", Proc. EUROSPEECH'91 - European Conference on Speech Technology, ESCA, Genoa, Italy, 1991.