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

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

Semi-Automatic Generation of Cloze Question Distractors Effect of Students' L1

Juan Pino, Maxine Eskenazi

Language Technologies Institute
Carnegie Mellon University
{jmpino,max}@cs.cmu.edu

Abstract

We describe a method to semi-automatically generate incorrect choices, or distractors, for cloze (fill-in-the-blank) questions. We generated distractors aimed at revealing what type of misunderstanding a student was having. English as a Second Language learners answered a series of cloze questions that presented distractors generated by our method. We analyzed their answers in order to see how native languages influence the type of distractor that is chosen. With this preliminary study, we intend to further individualize the use of an intelligent tutoring system for vocabulary learning.

1. Introduction

Our goal is to individualize the use of an intelligent tutor for English as a Second Language (ESL) vocabulary learning. This tutor, REAP, retrieves authentic documents from the web and filters them according to length, text quality, readability level [1] and topic in order to create a database of annotated documents. Annotations are used to provide students with documents that match their learning needs and interests. In a tutoring session, document reading is followed by a series of questions that both assess the students and require them to practice vocabulary. One of the question types used in this system is the cloze question type, an example of which is shown in Figure 1.

Select the word that best completes the sentence.
Memoranda, correspondence, and survey forms __ this display at the local history museum.

- ● comprise
- ○ intrinsic
- ○ trace
- ○ transfer
- ○ transform

Figure 1: Example of cloze question

Currently, students read documents containing target words that they need to learn. Cloze questions following the reading focus on the target words that appeared in the reading. However, two students seeing different readings with the same target word will see the same cloze question for this target word. We seek to individualize the questions by presenting distractor types, i.e. wrong choices, that are more likely to be picked and that indicate a specific deficiency or misunderstanding in vocabulary knowledge. In this paper, we present our method to generate different types of distractors automatically and in a preliminary study, we analyze which cloze questions, more specifically which distractor types, are chosen by students according to their native language.

2. Background

Goodrich [2] analyzed the potency (how attractive a distractor is) and discrimination power (how a distractor discriminates between high proficiency and low proficiency students) of eight distractor types that were generated manually; he also analyzed differences between populations with the same native language and from different geographical areas. Two of the distractor types–orthographic and morphological–are similar to the types that are generated in the present study. These distractor types were moderately potent: they were the 6^{th} and 5^{th} most potent distractors among the eight types considered. They were relatively discriminative as they were the 5^{th} and 3^{rd} most discriminative distractors among the eight types considered. Major differences between populations were not found. We picked these distractors in our study because they were suitable for automatic generation. Our analysis focuses on the differences between native languages with respect to five distractor types rather than on one unique native language. We investigate how attractive distractor types are with respect to different native languages.

Aldabe and colleagues [3] designed a system to automatically generate questions and distractors. Their questions focus on grammar and distractors and are produced by a morphological generator. They evaluated the quality of the generated distractors for error correction and multiple-choice question types [4] by presenting them to human experts. Our focus is on vocabulary rather than grammar; we also seek to match distractor types to native languages by presenting them to students rather than to human experts.

Pallier and colleagues [5] studied the influence of native-language phonology on lexical access and concluded that non-native speakers of a language might treat two words as homophones whereas natives would not. They conclude that the phonological representation of words depends on the native language. Weber and Cutler [6] showed that non-natives, more than natives, are distracted by pictures containing words with vowels related to the target in an eye-tracking experiment. This again tends to indicate that native language influences the phonological representation of foreign vocabulary. Heuven and colleagues [7] investigated how the number of orthographic neighbors, i.e. words differing by one letter, affects word recognition for native and non native speakers.

We next describe how we generate the distinct distractors used in our experiment, describe the setup of the experiment and discuss results in terms of distractor choice frequency and response time.

3. Distractor generation

We considered five different distractor types. In a preliminary study on spelling, we observed difference in performance depending on students' native languages. The types considered here assess spelling and other components of word knowledge. In an intelligent tutor, answers are recorded as an update of the student model; incorrect answers might indicate what specific problem a student has encountered. The first type (Morph) was a morphological variant of the target word. For example, if the target word was "bored", an incorrect answer would be "boring". In order to generate morphological variants, we used the XTAG system morphology database [8]. Several variant types were used, such as adding *-ing* or *-ed* to a verb, *-s* to a noun, *-er* or *-est* to an adjective. Gumnior and colleagues [9] showed that word production in a translation task is facilitated by morphological processing. In this experiment, words are not produced in isolation, instead they are shown in context. This distractor type was therefore designed to detect not only differences in morphological processing abilities but also in word integration skills. Fender [10] showed that Arabic speakers have better word integration skills than Japanese speakers. In this study, we compare Arabic and Chinese speakers. The second type (Orth) was an orthographic variant of the word. In order to generate this second distractor type, two or three consecutive letters of the target word were permuted. The resulting letter string was checked against a dictionary. We use the CMU Pronouncing Dictionary [11] which was also used to generate the phonetic distractors. The Orth distractor type was designed to pinpoint a difficulty with orthography. The third type of distractor (Phon) represented mapping from orthography to phonetics. In order to generate this type of distractor, two or three consecutive phonemes of the target word were permuted. More precisely, all words in the CMU Pronouncing Dictionary that had the same set of phonemes as the target word were considered and only those which had only two or three phonemes permuted were retained. The Phon type was designed to detect a deficiency in the phonetic knowledge of a word. The fourth type (OrthMorph) was a combination of the first and second types. The fifth type (PhonMorph) was a combination of the first and third types. The OrthMorph and PhonMorph types were designed to identify morphological, orthographic and phonetic knowledge. These types are further away from the target, therefore less plausible distractors: we expect them to be less often chosen than other types. Table 1 shows an example of each type of distractor.

Distractor Type	Target Word	Distractor
Morph	bored	boring
Orth	bread	beard
Phon	file	fly
OrthMorph	organ	groaning
PhonMorph	shared	shredded

Table 1: Examples for each Distractor Type

4. Experimental setup

In our experiment, we used 33 target words that ranged between rank 323 and rank 5886 of the lemmatized frequency list developed by Kilgarriff [12]. In our experiment, we used 33 target words that ranged between rank 323 and rank 5886 of the lemmatized frequency list developed by Kilgarriff [12]. By using this frequency range, we made sure that students would not know all words in the experiment. We first automatically selected words that produced both Orth and Phon distractors. There were 357 such words. Due to time constraints, we could only afford to use between 30 and 40 target words. Since the CMU pronouncing dictionary is quite comprehensive, target words could produce distractors that were rare words. We discarded those words. We also selected words that were suitable for ESL learners at an intermediate level. This manual selection retained 33 target words. For each of the 33 retained words, a cloze question was created. Each cloze question had one distractor of each of four of the five different types described previously. The distractors and the correct answer were displayed in random order.

Fifty-four students at an intermediate level of English, studying English as a Second Language at an American university participated in the experiment. The distribution of their native languages is displayed in Table 2. In our study, we had access to students mainly from two L1 populations: Arabic and Chinese. We concentrated on the two largest populations, comparing the answers of 22 Arabic speakers and 13 Chinese speakers.

Language	Number of students
Arabic	22
Chinese	13
Korean	7
Japanese	3
Others	9

Table 2: Distribution of native languages

5. Results

5.1. Overall performance

Table 3 compares the overall results for Arabic and Chinese speakers. Although the differences were not statistically significant, the Arabic speakers do better than Chinese speakers for words that they indicate that they are familiar with and they do worse on words they indicate that they do not know.

Arabic (with prior knowledge)	59.29%
Arabic (without prior knowledge)	31.49%
Chinese (with prior knowledge)	49.92%
Chinese (without prior knowledge)	43.25%

Table 3: Overall Performance for Chinese and Arabic Speakers

5.2. Distractor choice for Arabic and Chinese speakers

Figure 2 shows the distribution of distractor choice for Arabic and Chinese speakers when they have prior knowledge of the words. In general, Arabic and Chinese speakers choose correct answers significantly more often than distractors (p-values are respectively 2.2e-16 and 2.393e-09). The ranking of distractor choice frequency is the same for both Chinese and Arabic speakers. However, significance results differ, furthermore we noticed that the ranking is different when we do not control

for prior knowledge. For Chinese speakers, the Morph distractor type is chosen significantly (p = 0.005) more often than the Orth type. For Arabic speakers, this difference was marginally significant (p = 0.08). All remaining comparison of distractor types with adjacent rank do not give significant results. It is worth noticing that for Arabic speakers, the difference between the OrthMorph type and the PhonMorph is marginally significant (p = 0.09).

We also compared distractor choice between Arabic and Chinese speakers within each distractor type category. Significant differences were found between Arabic and Chinese speakers for morphological distractors. When prior knowledge was not controlled for, there was also a significant difference for the OrthMorph distractor type. Chinese speakers pick morphological distractors significantly more often than Arabic speakers. This is not surprising given that Chinese has hardly any morphology whereas Arabic has a rich morphology. On the other hand, Arabic speakers pick OrthMorph distractors significantly more often than Chinese speakers (with no controlled prior knowledge). An interpretation of this result is not as straightforward. This could be due to the fact that in Arabic, morphological derivation alters the vowels of stem and adds morphemes around the altered stem (called root and pattern morphology).

Figure 3 shows the distribution of distractor choice when the students do not have prior knowledge of the words. This time, we notice that the ranking distractor choice frequency is different than with prior knowledge and that the ranking is different between Chinese and Arabic speakers. Again, for both categories, correct answers were chosen significantly more often than Phon, the most popular distractor (p-values are respectively 0.0193 and 0.0025). When comparing the language groups within one distractor type, we find a significant difference for the OrthMorph distractor type (p = 0.0410).

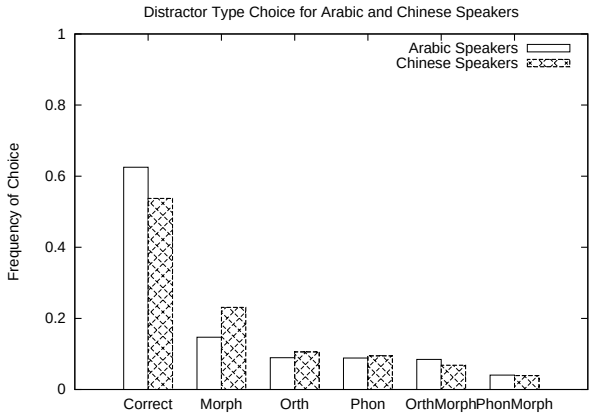

Figure 2: Comparison of Distractor choice for Arabic and Chinese speakers with prior knowledge

5.3. Timing

We also analyzed the response time of the participants in Figures 4 and 5. For both Arabic and Chinese speakers, distractor type ranking according to response time was different from the ranking according to choice frequency. When Arabic and Chinese speakers answered correctly, it took them less time on average than when they chose one of the distractors (except for

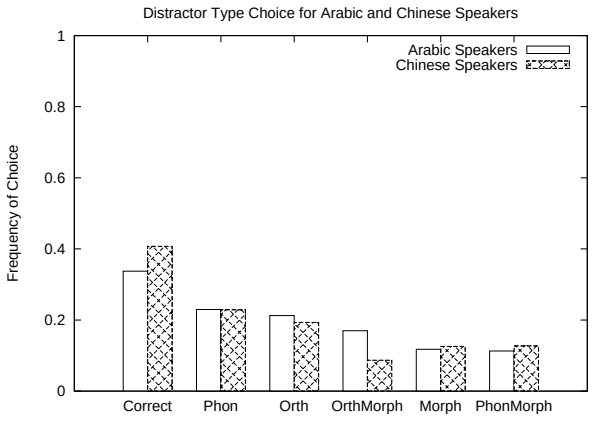

Figure 3: Comparison of Distractor choice for Arabic and Chinese speakers without prior knowledge

the OrthMorph type for Chinese speakers). This seems to indicate that the participants were careful at their task: if they were not confident about an answer, they took more time instead of answering quickly and randomly.

For Chinese speakers, the Orth distractor type had a substantially longer response time than other distractor types, namely the Phon type distractor. This is surprising given than the Orth and Phon were at the same level of choice frequency. This might indicate that with respect to response time, the Orth type is less attractive than the Phon type because it is chosen after a longer time.

When comparing Arabic and Chinese speakers' response time for each distractor type (Figure 4), we notice important differences for Phon distractors and OrthMorph distractors (about 10 seconds). The time difference for the OrthMorph corresponds to a higher frequency of OrthMorph type choice by Arabic speakers. However, the time difference for the Phon distractor corresponds to a lower frequency of Phon type choice by Arabic speakers. These observations do not allow to conclude to a relationship between distractor choice frequency and response time.

In the case of no prior knowledge, we observe important difference in response time between the two language groups for PhonMorph and for Orth distractor types. Again, the ranking of distractor types according to response time is different from the ranking according to choice frequency.

6. Conclusion and future work

In this study, we have described a method to automatically generate distractors for cloze questions; we also have shown differences in distractor choice that appear to be influenced by the native languages of the students. In our tutor, teachers are able to see students' results and answers. If they notice that a certain distractor type is chosen more often than other types, they can adapt their teaching in order to address this difficulty.

In the future, we would like to further individualize and automate the process of distractor selection. The initialization of this process would be based on the results of this study. However, the distractor selection process would dynamically evolve rather than following always the same policy. Finally, we would like to investigate the use of other distractor types, for example

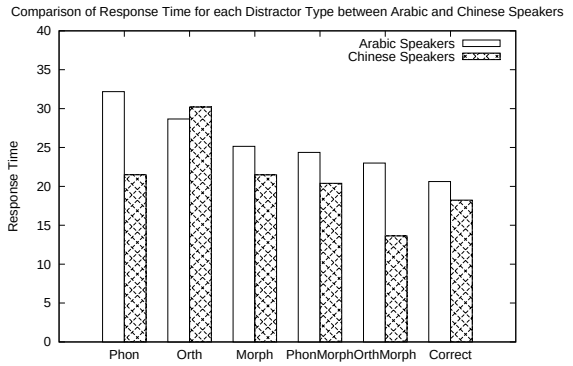

Figure 4: Response Time for Arabic and Chinese Speakers with Prior Knowledge for each Distractor Type

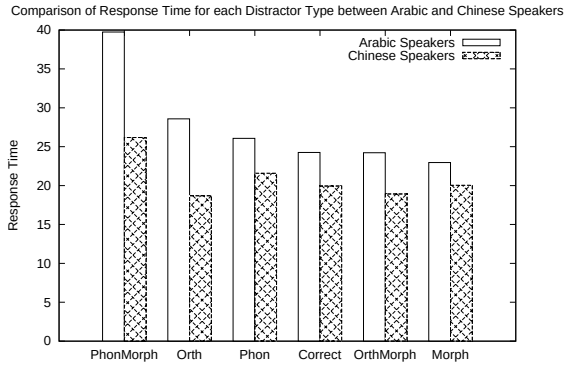

Figure 5: Response Time for Arabic and Chinese Speakers without Prior Knowledge for each Distractor Type

false cognates which are by definition distractors only for specific native languages.

7. Acknowledgments

We would like to thank Alan Juffs for making the study possible and Sharon Rosenfeld for reviewing this paper. This research is supported by NSF grant SBE-0354420. Any opinions, findings, conclusions or recommendations expressed in this material are the authors, and do not necessarily reflect those of the sponsor.

8. References

[1] K. Collins-Thompson and J. Callan, "Predicting reading difficulty with statistical language models," *Journal of the American Society for Information Science and Technology*, vol. 56, no. 13, pp. 1448–1462, 2005.

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

[3] I. Aldabe, M. Lopez de Lacalle, M. Maritxalar, E. Martinez, and L. Uria, *Arikiturri: an Automatic Question Generator Based on Corpora and NLP techniques*, ser. Lecture Notes in computer science. Springer, 2006, vol. 4053, pp. 584–594.

[4] I. Aldabe, M. Maritxalar, and E. Martinez, "Evaluating and improving the distractor-generating heuristics," in *Workshop on NLP for Educational Resources. In conjuction with RANLP07*, 2007, pp. 7–13.

[5] C. Pallier, A. Colomé, and N. Sebastián-Gallés, "The influence of native-language phonology on lexical access: exemplar-based vs. abstract lexical entries," *Psychological Science*, vol. 12, pp. 445–449, 2001.

[6] A. Weber and A. Cutler, "Lexical competition in non-native spoken-word recognition," *Journal of Memory and Language*, vol. 50, no. 1, pp. 1–25, 2004.

[7] W. J. Van Heuven, T. Dijkstra, and J. Grainger, "Orthographic neighborhood effects in bilingual word recognition," *Journal of Memory and Language*, vol. 39, no. 3, pp. 458–483, 1998.

[8] C. Doran, D. Egedi, B. A. Hockey, B. Srinivas, and M. Zaidel, "Xtag system – a wide coverage grammar for english," in *Proceedings of the 15th International Conference on Computational Linguistics*, vol. 2, 1994.

[9] H. Gumnior, J. Bölte, and P. Zwitserlood, "A chatterbox is a box: Morphology in german word production," *Language and Cognitive Processes*, vol. 21, no. 7, pp. 920–944, 2006.

[10] M. Fender, "English word recognition and word integration skills of native arabic- and japanese-speaking learners of english as a second language," *Applied Psycholinguistics*, vol. 24, no. 02, pp. 289–315, 2003.

[11] "Carnegie mellon pronouncing dictionary," http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

[12] A. Kilgarriff, "Putting frequencies in the dictionary," *International Journal of Lexicography*, vol. 10, no. 2, pp. 135–155, 1997.