Evaluating Human and Automated Generation of Distractors for Diagnostic Multiple-Choice Cloze Questions to Assess Children's Reading Comprehension

Yi-Ting Huang[1, 2] and Jack Mostow[2]

[1] National Taiwan University, Taipei, Taiwan
[2] Carnegie Mellon University, Pittsburgh, PA, United States of America
d97008@im.ntu.edu.tw, mostow@cs.cmu.edu

Abstract. We report an experiment to evaluate DQGen's performance in generating three types of distractors for diagnostic multiple-choice cloze (fill-in-the-blank) questions to assess children's reading comprehension processes. Ungrammatical distractors test syntax, nonsensical distractors test semantics, and locally plausible distractors test inter-sentential processing. 27 knowledgeable humans rated candidate answers as correct, plausible, nonsensical, or ungrammatical without knowing their intended type or whether they were generated by DQGen, written by other humans, or correct. Surprisingly, DQGen did significantly better than humans at generating ungrammatical distractors and slightly better than them at generating nonsensical distractors, albeit worse at generating plausible distractors. Vetting its output and writing distractors only when necessary would take half as long as writing them all, and improve their quality.

Keywords: Question generation, reading comprehension, cloze, distractors

1 Introduction

Traditionally, generation of questions to assess reading comprehension relied on humans – either teachers (and students) during instruction, or materials developers beforehand. More recently, the less labor-intensive approach of automated question generation has been used for multiple tasks, such as inserting comprehension checks in a reading tutor [1], generating comprehension instruction [2], testing vocabulary [3], recognizing children's spoken questions [4], assessing closed-domain knowledge [5, 6], evaluating language proficiency [7-10], and assisting academic writing [11].

One type of question especially conducive to automated generation is the multiple choice cloze (fill-in-the-blank) question, in which one word in a sentence is replaced with a blank. Answering without guessing requires having relevant background knowledge and understanding the context in order to select the best word from a list of options for completing the sentence. Cloze questions are used in many standardized tests, such as SAT (Scholastic Aptitude Test), TOEFL (Test of English as a Foreign Language), and TOEIC (Test of English for International Communication). Research has explored automated generation of cloze questions for various purposes, for example to test comprehension of important concepts in textbooks [5]. In the domain of

adfa, p. 1, 2011.

language learning, a growing number of studies explain how to generate such questions to test English language proficiency with verbs [7], prepositions [8], adjectives [9], and grammar patterns [10]. Especially in language learning, cloze questions can test the ability to decide which word is consistent with the surrounding context. Thus they tap comprehension processes that judge various types of consistency, such as syntactic, semantic, and inter-sentential, in the course of constructing a situation model that represents "the content or microworld that the text is about" [12]. In brief, these processes encode sentences, integrate them into an overall representation of meaning, notice gaps and inconsistencies, and repair them [13, 14].

DQGen (Diagnostic Question Generator) [15] generates cloze questions for diagnostic assessment of a child's comprehension while reading a given text. As Fig. 1 illustrates, DQGen's questions have four components. The *stem* is the truncated sentence, "That helps your body find and kill _____." The *context* is the text preceding the stem. The *correct answer* is by definition the deleted original word "germs." The *distractors* are the other candidate completions.

Some of those cells patrol your body. They are hungry, and they eat germs! Some stop the trouble germs make. Others make antibodies. They stick to germs. That helps your body find and kill _____.

1. are – **ungrammatical**
2. intestines – **nonsensical** (but grammatical)
3. terrorists – **plausible** (meaningful by itself but incorrect given the preceding text)
4. germs – **correct**

Fig. 1. Annotated example of a multiple-choice cloze question generated by DQGen

To detect failures in different comprehension processes, DQGen uses three types of distractors. Each type of distractor indicates a different type of comprehension failure when chosen by a child instead of the correct answer. DQGen classifies the word "are" as ungrammatical because it has the wrong part of speech. DQGen classifies "intestines" as nonsensical because "find and kill intestines ." does not occur in the Google N-grams corpus. DQGen classifies "terrorists" as plausible only locally because "find and kill terrorists ." occurs in the Google N-grams corpus, but "terrorists" is topically unrelated to the preceding paragraph. DQGen classifies "germs" as correct because it was the last word of the original sentence. Aggregating children's performance over questions with these three types of distractors should not only assess their comprehension, but profile the difficulties faced by a given child or posed by a given text. For instance, a child who processes syntax and semantics but not the relation of a sentence to the context preceding it would reject the ungrammatical and nonsensical distractors, but pick the plausible distractor as often as the correct answer.

DQGen uses a generate-and-test approach. It chooses a candidate at random from a source of candidates for that type of distractor, and rejects the candidate if it does not satisfy the constraints for that type, e.g. that ungrammatical distractors must have the wrong part of speech. Mostow and Jang [15] evaluated DQGen by itself; here we evaluate the current (2014) version of DQGen against human performance. Section 2 describes our experiment. Section 3 reports results. Section 4 concludes.

2 Experimental design

To evaluate DQGen against human performance, we had to specify the task being performed and the criteria by which to evaluate it. Given a text with some sentences selected to turn into stems by deleting the last word, the task was to generate a distractor of each type – ungrammatical, nonsensical, and plausible.

Our principal evaluation criterion was whether a generated distractor achieved its purpose according to human judges blind to its source (DQGen or human), its intended type (ungrammatical, nonsensical, or plausible), and the correct (original) answer. An additional evaluation criterion was time: we wanted to know how long it took humans to rate or write each type of distractor. Besides quantifying the relative difficulty of rating vs. writing the three types of distractors, the practical purpose of this information was to predict which would be faster – writing distractors by hand, or hand-vetting distractors generated by DQGen.

MATERIALS: To enable controlled evaluation of distractors, we gave DQGen and humans the same 7 texts from Project LISTEN's Reading Tutor [16], containing a total of 16 cloze stems and chosen to ensure that DQGen could generate each type of distractor for each stem.

APPARATUS: To run the experiment, we implemented a website in PHP and connected it to a MySQL database server that logged a timestamped event for each page entrance or exit, keyboard input, or menu selection. The database also kept track of each participant's position in the protocol in order to continue at the same point after an interruption, and to avoid repeating any of the protocol.

PARTICIPANTS: To recruit human experts proficient in English and sufficiently knowledgeable about reading comprehension to rate and write distractors, we posted a request to Carnegie Mellon's doctoral Program in Interdisciplinary Research (www.cmu.edu/pier) and to the Society for the Scientific Study of Reading (triplesr.org). The request directed participants to the website for the experiment.

After data cleaning to filter out data from in-house software testing, failed attempts to log in, unfinished protocols, and two null ratings, we had data for 27 participants.

PROCEDURE: The experimental protocol consisted of logging into the experiment website, a brief introduction, the two main tasks (first rating, then writing), and finally a survey with a series of optional typed-input questions about various aspects of the experiment.

The introduction thanked participants for "helping our research by doing two tasks: rating (the first task) and designing (the second task) multiple choice cloze (fill-in-the-blank) items to assess children's reading comprehension." It explained that in the first task, they would read texts containing a total of 8 cloze stems, see different candidate completions of each stem, and classify each completion as **Correct**, **Plausible**, N**onsensical**, or **Ungrammatical.** It showed the annotated example in Fig. 1 and:

> Please classify each choice on its own merits, independently of the others.
> Your responses will be timed as a measure of the effort they require.
> Therefore you will not get an opportunity to revise them.
> Also, please try to avoid interruptions during a text.
> However, pausing between texts is fine.

In the rating task, participants read 3-4 texts containing a total of 8 stems. Stems appeared on a new screen with this note: "If you need to reread the text first, please click on the *Previous* button above. Otherwise, click on one of the 4 buttons below to classify the following completion (independently of the others)." The button for each rating included its description shown in Fig. 1. Participants rated seven candidate single-word completions, one at a time, for each stem, e.g., "*The next morning, Silly Pilly was ready to go to ____.*" The seven candidates, reordered randomly for each participant, consisted of the correct answer ("*school*"), the three distractors generated by DQGen ("*along*", "*slang*", and "*breakfast*"), and three authored by humans (e.g. "*blue*", "*slip*", and "*home*"). The writing task was similar:

> In the second task, you will read texts that contain cloze items. You will be prompted to type in four **1-word** completions of each cloze item, one completion of each kind. **These words should be no harder for a child than the reading level of the text.**

To avoid problematic input such as null responses, typos, and non-words, we included code to reject them and prompt for a replacement, but these events, averaging 25 seconds, occurred for only 11 of the 504 human-written distractors in our data.

ASSIGNMENT TO CONDITIONS: All participants did rating before writing, which we considered harder and in fact averaged about 3 times as long per completion. To avoid text-specific bias, we counter-balanced the study design so that half the participants (the "AB" group) rated completions for the 8 stems in set A and then wrote completions for the 8 stems in set B, and the other half (the "BA" group) rated completions for the stems in set B and then wrote completions for the stems in set A.

Participants in each group rated the same distractors generated by DQGen, but they rated different distractors authored by humans, so as to give us a more diverse sample. To limit the protocol duration, each participant rated distractors authored by only one participant from the other group. Accordingly, we used the following algorithm to assign participants to rate human-authored distractors.

The first participants saw distractors written by staff experienced with cloze questions. However, as soon as participants completed the protocol, the distractors written by these protocol completers became available for subsequent participants to rate. Once a participant completed the protocol by writing distractors for set B, another participant was assigned to rate them, and to write distractors for set A. Similarly, those distractors were eventually (if ever) rated by some subsequent participant assigned to rate set A and write distractors for set B, and so on.

This "daisy-chaining" algorithm assigned each new participant to rate cloze items from whichever set (A or B) had been rated so far by fewer participants who had finished the protocol. It chose human-authored distractors not yet rated by anyone who had finished the protocol. Consequently, all 27 participants rated distractors generated by DQGen for either set A or set B. 21 participants' distractors got rated – 16 participants with one rating per distractor, and the other five with two. Our data set contains no ratings for the remaining six participants' distractors, either because nobody rated them, or because we discarded data from other participants who may have rated them but didn't finish the rest of the protocol.

3 Results

Table 1 shows the percentage of ratings of each intended distractor type as Ungrammatical, Nonsensical, Plausible, or Correct, based on 1486 ratings by 27 raters of 16 correct answers, 48 distractors generated by DQGen, and 504 distractors written by 21 humans. Inter-rater reliability was substantial on distractors generated by DQGen (Fleiss' Kappa = 0.66). Only 40 human-authored distractors were rated by more than one rater, namely 5 sets of 8 distractors rated by a pair of raters. Cohen's Kappa for each pair of raters averaged 0.46 (N = 5, SD 0.25), i.e., only moderate agreement, vs. 0.60 (SD 0.09), close to substantial agreement, on the 8 DQGen-generated distractors they both rated, but the two means did not differ reliably on a paired T-test ($p=0.31$).

Table 1. Confusion matrix for ratings of DQGen's and human distractors and correct answers

Rating:	**Ungrammatical**		**Nonsensical**		**Plausible**		**Correct**	
Intended type:	DQGen	Human	DQGen	Human	DQGen	Human	DQGen	Human
Ungrammatical	**93% > 81%** [a]		4%	16%	3%	1%	0%	2%
Nonsensical	14%	5%	**81% > 74%** [b]		5%	20%	0%	1%
Plausible	2%	2%	23%	23%	**54% < 63%** [c]		21%	13%
Correct	0%		2%		18%		80%	

a. Chi-square $p < 0.001$; b. $p = 0.089$; c. $p = 0.053$

The **boldfaced** diagonal entries in Table 1 compare the percentages of ratings that agreed with the intended types of DQGen- and human-generated distractors. To determine which differences were not only reliable but likely to generalize to unseen data from similar cloze stems and raters, we used a logistic mixed-effects model. Like logistic regression, it predicted a binary outcome – whether the rating of a distractor will agree with its intended type – as the log odds ratio of the probability of agreement over the probability of disagreement. It used random effects to model variation in cloze stems or raters. To find the model that fit the data best, we used backward model selection, starting with five predictors we expected could affect the outcome. Three were fixed effects: distractor source, intended type, and their interaction. Two were random effects: stem and rater (just their intercepts, not their slopes, which our data was too sparse to estimate). We kept removing the weakest predictor (the one with the highest p-value) until doing so stopped improving model fit in a Likelihood Ratio Test. We now relate the resulting model in Table 2 to the ratings in Table 1:

Main effect of intended distractor type: Compared to their nonsensical distractors, both DQGen and humans generated significantly worse ($p < 0.02$) plausible distractors, with a trend ($p < 0.1$) toward better ungrammatical distractors.

No main effect of source: Surprisingly, DQGen's distractor quality did not differ significantly overall from humans'.

Interaction of source with distractor type: Although DQGen and humans did not differ significantly overall, they differed for some distractor types after adjusting for the fixed effect of distractor type and the random effect of stem. DQGen's ungrammatical distractors were significantly ($p < 0.001$) better than humans', its plausible distractors were probably ($p \sim 0.05$) worse than humans', and there was a trend ($p < 0.1$) for its nonsensical distractors to be better than humans'.

No random effect of individual rater: We would have expected a rater effect if some raters were systematically worse, e.g., rated at random. The absence of such an effect reassuringly suggests the results are likely to generalize to future similar raters.

Random effect of stem: Performance differed reliably by stem (SD = 0.35), i.e., the best-fitting model had a (1 | stem) ~ $N(0, 0.35^2)$ distribution of random per-stem intercepts. For some stems, raters could not tell correct answers from plausible distractors, as the error analysis in Section 3.2 below will discuss further.

Table 2. Best-fitting model of agreement; reference base for distractor_type is Nonsensical.

Model: *agreement ~ distractor_type + source × distractor_type + (1 | stem)*		
Random effects:	**Variance:**	**SD:**
stem	0.12	0.35
Fixed effects:	**β coefficient:**	***p*-value:**
intercept	1.08	<0.001
distractor_type = Ungrammatical	0.40	0.098
distractor_type = Plausible	-0.53	0.014
distractor_type = Ungrammatical × source = DQGen	1.11	<0.001
distractor_type = Nonsensical × source = DQGen	0.41	0.082
distractor_type = Plausible × source = DQGen	-0.39	0.053

TIME ANALYSIS: To see whether DQGen could speed up human authoring, we compared the time for humans to rate versus write distractors. They averaged about 5 seconds to rate a choice and about 19 seconds to write any type of distractor. Based on Table 1, rating a distractor generated by DQGen and rewriting it only if unacceptable would average (5 seconds) + (1 – agreement rate) × (19 seconds) = about 10 seconds, barely half of the 19 seconds to write it by hand. Moreover, 92% of distractors would match their intended type if vetting is perfect, i.e. rates DQGen-generated distractors properly by definition. Only 73% of human-authored distractors do so.

To analyze effects on the time to rate a choice, we used mixed-effects linear regression starting with source, type, rating agreement and their interaction as fixed effects, and stem and rater as random effects. Backward model selection led to the model in Table 3:

No main effects: Rating time didn't differ reliably by source, type, or agreement.

Random effects: Rating time differed reliably by both stem and rater.

Interaction of agreement with intended type (p < 0.001)**:** Rating was significantly faster when it agreed with choices intended to be correct (4.6 s < 8.7 s), ungrammatical (4.3 s < 7.2 s), or nonsensical (5.2 s < 5.7 s). For plausible distractors, rating was slower (6.6 s > 5.8 s) (albeit not significantly) when it agreed with intended type, perhaps because confirming that a distractor is plausible requires additional thought.

Table 3. Best-fitting model of time to rate a choice

Model: *duration ~ intended_type × agreement + (1 | stem) + (1 | rater)*		
Random effects:	**Variance:**	**SD:**
stem	3.48	1.87
rater	1.92	1.38
Fixed effects:	**β coefficient:**	***t*-value:**
intercept	6.26	10.32
intended_type = Ungrammatical × agreement = agree	-1.89	-4.76
intended_type = Nonsensical × agreement = agree	-0.97	-2.37
intended_type = Plausible × agreement = agree	0.14	0.33
intended_type = Correct × agreement = agree	-1.61	-3.24

ERROR ANALYSIS: To shed light on which distractors were rated differently than their intended type, and why, we now discuss the off-diagonal cases in Table 1, most frequent first:

Distractors intended to be plausible but rated as nonsensical: One possibility is that the raters disregarded "meaningful by itself" in our definition of plausible and took context into account in rating some distractors as nonsensical.

Distractors intended to be plausible but rated as correct, or vice versa: For instance, raters performed below or near chance on two sentences from a speech by Bill Clinton where his actual word fit no better than the plausible distractor:

- our people have always mustered the determination to construct from these crises the pillars of our ____. [history]

Only 21% of the ratings of "history" classified it as correct, vs. 86% for "democracy."

- Clearly America must continue to lead the world we did so much to __. [make]

Only half of the ratings for "make" classified it as correct.

DQGen assumes that the correct answer fits better than topically unrelated distractors. This assumption fails when lack of topicality fails to disqualify a plausible distractor.

Distractors generated by DQGen to be nonsensical, but rated as ungrammatical: For instance, 13 of 14 raters classified the distractor "share" as ungrammatical here:

- We nip if they stray too far from ____. [home]

DQGen chooses nonsensical distractors to have the same part of speech as the correct answer, in this case the noun “home”. The word “share” can be a noun, but evidently raters perceived it here as a verb and hence ungrammatical.

Distractors written by humans to be ungrammatical, but rated as nonsensical: 6 of 13 raters classified “brave,” “flows,” “light,” “politics,” or “run” as nonsensical in:

- Now, the sights and sounds of this ceremony are broadcast instantaneously to billions around the ____. [world]

Perhaps the raters parsed them as nouns, but their authors did not (except “politics”). In generating ungrammatical distractors, DQGen considers alternative parts of speech.

Distractors written by humans to be nonsensical, but rated as plausible: For instance, of the supposedly nonsensical distractors written by humans for the sentence “We nip if they stray too far from ____.”, “Bananas,” “beaches,” “beans,” “England,” “heaven,” “muscle,” “pizza, ” and “sheep” were indeed classified as nonsensical (each by a different rater), but “England,” “heaven, “rivers”, “sheep,” and “water” were classified as plausible (each by some other rater). The last three distractors could be attributed to authors taking context into account, and therefore considering them nonsensical even though they’re plausible out of context. Apparently they were unable to disregard context in deciding whether a word is nonsensical. The disagreement on “England” and “heaven” suggests that it’s less clear-cut how to rate them, or perhaps that their plausibility or lack thereof depends on the extent to which the rater ignores context. Evidently writing or judging nonsensical distractors is a difficult task for human raters who know the context, because they have trouble disregarding it. Depriving humans of the context would make both tasks easier for humans. In contrast, DQGen by its very design disregards the context when generating nonsensical or ungrammatical distractors.

4 Conclusion

This paper contributes to automated diagnostic assessment of children’s reading comprehension by comparing the 2014 version of DQGen against human performance in generating ungrammatical, nonsensical, and plausible distractors for cloze stems. We had assumed that human performance was a gold standard to aspire to, and an existence proof of the level of performance possible; the gap would show where further progress was possible and needed.

Surprisingly, DQGen did not differ significantly overall from human performance, and actually beat humans at generating ungrammatical and nonsensical distractors. Its plausible distractors were too plausible, i.e. rated as correct answers 18% of the time. Error analysis elucidated the performance differences: DQGen considers all parts of speech and distinguishes local from contextual plausibility, but needs stronger heuristics than topicality to reliably generate distractors implausible in the larger context. We projected that vetting DQGen’s output and writing distractors only if needed, rather than writing them all, would take only half the time and yield better distractors.

Previous evaluations of automatically generated cloze questions relied on expert critiques or crowdsourced human performance at answering them. We found one study [5] in which three experts compared their estimated time to write cloze questions for different texts with vs. without automated assistance. Our study was much more tightly controlled, evaluating DQGen-generated vs. human-authored distractors for the same 16 cloze stems by what percentage of ratings by knowledgeable judges agreed with the distractors' intended types, and by the exact logged time to write or rate them. To gauge the generalizability of our results to similar stems and raters, we used mixed-effects models to analyze 1486 ratings by 27 raters of 16 correct answers, 48 distractors generated by DQGen, and 504 distractors authored by 21 humans.

Small-scale crowd-sourced expert rating and writing of distractors enabled controlled comparison of both the quality of each type of distractor and the time to write or rate them. It exposed the influence of the preceding text on raters' ability to distinguish nonsensical from plausible distractors. Future studies should eliminate this influence by having humans rate or write distractors for cloze stems before seeing their context, and only then decide which plausible distractors do not fit the context.

Limitations of this study leave ample room for future work. We used 16 cloze stems from just seven stories. To enable controlled comparison of distractors, we took these stems as givens, so we did not evaluate the percentage of sentences turned into cloze stems (yield), and we evaluated just the distractors, not the cloze stems, nor the overall quality of the resulting questions in diagnostic comprehension assessment. We used education and reading researchers blind to intended type to rate distractors, rather than validate them against expert diagnostic assessments of children. Finally, distractors based on deeper models of comprehension processes such as inter-sentential inference may enable more reliable and informative diagnostic assessments.

Acknowledgements: This research was funded in part by the National Science Foundation through Grant IIS1124240, and by the Taiwan National Science Council's Graduate Students Study Abroad Program. We thank everyone who rated and wrote distractors; reviewers and H. Jang for their comments; and Prof. Y. S. Sun at National Taiwan University and Dr. M. C. Chen at Academia Sinica for supporting the first author. The opinions expressed are those of the authors and do not necessarily represent the views of the National Science Foundation or the National Science Council.

References

1. Mostow, J., J.E. Beck, J. Bey, A. Cuneo, J. Sison, B. Tobin, and J. Valeri. Using automated questions to assess reading comprehension, vocabulary, and effects of tutorial interventions. *Technology, Instruction, Cognition and Learning*, 2004. *2*(1-2): p. 97-134.
2. Mostow, J. and W. Chen. Generating Instruction Automatically for the Reading Strategy of Self-Questioning. *Proceedings of the 14th International Conference on Artificial Intelligence in Education*, 465--472. 2009. Brighton, UK. IOS Press.
3. Gates, D., G. Aist, J. Mostow, M. Mckeown, and J. Bey. How to Generate Cloze Questions from Definitions: a Syntactic Approach. *Proceedings of the AAAI Symposium on Question Generation* 2011. Arlington, VA. AAAI Press.

4. Chen, W., J. Mostow, and G.S. Aist. Recognizing Young Readers' Spoken Questions. *International Journal of Artificial Intelligence in Education*, 2013. *21*(4): p. 255-269.
5. Mitkov, R., L.A. Ha, and N. Karamanis. A computer-aided environment for generating multiple choice test items. *Natural Language Engineering*, 2006. *12*(2): p. 177-194.
6. Agarwal, M. and P. Mannem. Automatic gap-fill question generation from text books. *Proceedings of the 6th Workshop on Innovative Use of NLP for Building Educational Applications*, 56-64. 2011Association for Computational Linguistics.
7. Sumita, E., F. Sugaya, and S. Yamamoto. Measuring non-native speakers' proficiency of English by using a test with automatically-generated fill-in-the-blank questions. In *Proceedings of the Second Workshop on Building Educational Applications Using NLP*. 2005, Association for Computational Linguistics: Ann Arbor, Michigan, p. 61-68.
8. Lee, J. and S. Seneff. Automatic generation of cloze items for prepositions. *INTERSPEECH*, 2173-2176. 2007.
9. Lin, Y.-C., L.-C. Sung, and M.C. Chen. An automatic multiple-choice question generation scheme for english adjective understanding. *Workshop on Modeling, Management and Generation of Problems/Questions in eLearning, the 15th International Conference on Computers in Education (ICCE 2007),* 137-142. 2007.
10. Huang, Y.-T., M.C. Chen, and Y.S. Sun. Personalized Automatic Quiz Generation Based on Proficiency Level Estimation. *20th International Conference on Computers in Education (ICCE 2012)* 2012. Singapore.
11. Ming, L., R.A. Calvo, A. Aditomo, and L.A. Pizzato. Using Wikipedia and Conceptual Graph Structures to Generate Questions for Academic Writing Support. *Learning Technologies, IEEE Transactions on*, 2012. *5*(3): p. 251-263.
12. Graesser, A.C. and E.L. Bertus. The Construction of Causal Inferences While Reading Expository Texts on Science and Technology. *Scientific Studies of Reading*, 1998. *2*(3): p. 247-269.
13. van den Broek, P., M. Everson, S. Virtue, Y. Sung, and Y. Tzeng. Comprehension and memory of science texts: Inferential processes and the construction of a mental representation. In J.L. J. Otero, & A. C. Graesser, Editor, *The psychology of science text comprehension*. Erlbaum: Mahwah, NJ, 2002.
14. Kintsch, W. An Overview of Top-Down and Bottom-Up Effects in Comprehension: The CI Perspective. *Discourse Processes A Multidisciplinary Journal*, 2005. *39*(2&3): p. 125-128.
15. Mostow, J. and H. Jang. Generating Diagnostic Multiple Choice Comprehension Cloze Questions. *NAACL-HLT 2012 7th Workshop on Innovative Use of NLP for Building Educational Applications*, 136-146. 2012. Montréal. Association for Computational Linguistics.
16. Mostow, J. Lessons from Project LISTEN: What Have We Learned from a Reading Tutor that Listens? (Keynote). *Proceedings of the 16th International Conference on Artificial Intelligence in Education*, 557-558. 2013. Memphis, TN. Springer.