Automatic Generation of Context-Based Fill-in-the-Blank Exercises Using Co-occurrence Likelihoods and Google *n*-grams

Jennifer Hill and **Rahul Simha**
The George Washington University
800 22nd Street NW
Washington, DC 20052, USA
jenhill@gwu.edu
simha@gwu.edu

Abstract

In this paper, we propose a method of automatically generating multiple-choice fill-in-the-blank exercises from existing text passages that challenge a reader's comprehension skills and contextual awareness. We use a unique application of word co-occurrence likelihoods and the Google n-grams corpus to select words with strong contextual links to their surrounding text, and to generate distractors that make sense only in an isolated narrow context and not in the full context of the passage. Results show that our method is successful at generating questions with distractors that are semantically consistent in a narrow context but inconsistent given the full text, with larger n-grams yielding significantly better results.

1 Introduction

According to the American Library Association, approximately 43% of Americans have reading skills at or below the most basic level of prose literacy, defined as the ability to "search, comprehend, and use information from continuous texts" (Baer et al., 2009). These underdeveloped literacy skills are in many cases the result of poor reading comprehension. Results from a large-scale national survey indicate that most adult learners with low literacy skills have "difficulty integrating and synthesizing information from any but the simplest texts," likely due to a number of factors including poor phonemic awareness, vocabulary understanding, and reading fluency (Krudenier, 2002). It also suggests that adults in basic education programs are more likely to view reading as simply a decoding task rather than a multifaceted skill involving semantic processing, active memory, and inference.

One method of addressing weak reading skills is the cloze, or fill-in-the-blank (FITB), exercise. These exercises involve strategically removing *target words* from a text and requiring the reader to identify the missing word among a list of *distractors*. However, while FITB exercises can be valuable resources for practicing and improving reading skills, they are time consuming to create by hand. The goal of this paper is to suggest a method for automatically creating such exercises from existing text passages. Such a method would allow for significantly faster and less costly exercise creation on a larger scale, and would allow for nearly any desired reading materials to serve as a learning resource.

We propose that, for native English speakers, a good reading comprehension question challenges the reader not with syntactic errors or unusual word senses, but rather with contextual inconsistencies. Figure 1 gives an example of the type of question we wish to generate: when looking at a narrow context, all four of the word choices are logical selections for the blank, but when the meaning implied by the surrounding text is taken into account, only one choice is sensible. This type of exercise encourages engagement and focus while reading: as a well-formed question should not have obvious inconsistencies within a narrow reference frame, a reader must actively construct meaning as they read in order to identify the correct answer.

In this paper, we propose a method of automatically generating FITB questions from existing text

Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications, pages 23–30,
San Diego, California, June 16, 2016.

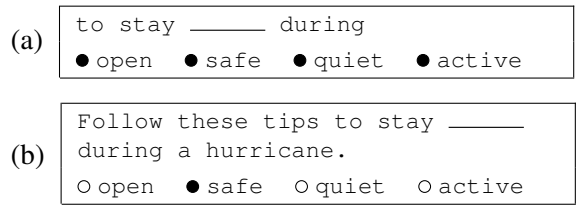

Figure 1: An example illustrating the premise behind our exercises: (a) In a narrow context, all four word choices are equally fitting; (b) In the full context, only the target word logically fits

passages that follow this context-specific pattern, using a unique application of word co-occurrence likelihoods and the Google Books n-gram corpus (Michel et al., 2010).

2 Previous Work

Ours is not the first paper to address the task of generating fill-in-the-blank questions. Many previous studies focus on automatically creating exercises specifically for language learning and vocabulary assessment. Sakaguchi et al. (2013) describe a method of generating distractors for assessing an ESL reader's ability to distinguish semantic nuances between vocabulary words. Brown et al. (2005) utilize WordNet word relations and frequencies to generate distractors for vocabulary words from equally-challenging terms. Pino and Eskenazi (2009) and Goto et al. (2010) both explore different methods of generating distractors of different classes designed to indicate particular deficiencies in phonetic or morphological vocabulary mastery.

Others focus on generating exercises for quizzing or knowledge testing purposes. Agarwal and Mannem (2011) explore generating gap-fill exercises from informative sentences in textbooks, while Karamanis and Mitkov (2006) locate suitable distractors for medical texts from domain-specific documents. Both of these methods choose distractors from other sentences in a constrained set of source texts rather than relying on external corpora.

A few studies have focused on more comprehension-specific exercises, generating distractors that are semantically similar to the target word. Zesch and Melamud (2014) propose a method of generating semantically similar distractors to the target word using context-sensitive lexical inference rules. The distractors generated using this method are contextually and semantically similar to the target word, but not in the context being used in the sentence. Kumar et al.'s RevUP system (2015) utilizes a word vector model trained on the desired text domain to find semantically-similar words and verifies their similarity using WordNet synsets. Aldabe et al. (2009) generate semantically-similar distractors using distributional data obtained from the British National Corpus, and also, like our study, utilize the Google n-grams corpus to determine each generated distractor's probability of occurring with its surrounding terms. However, their study differs from ours in that they utilize the Google n-grams solely for validating that their chosen distractors make sense, whereas we use the corpus for the actual generation of the distractors.

Perhaps the closest cousin to our proposed method can be found in the DQGen system (Mostow and Jang, 2012). DQGen generates cloze questions designed to test different types of comprehension failure in children, one of which involves creating "plausible" distractors that create contextually sensible sentences in isolation but do not fit in the context of the rest of the text. Their system also utilizes the Google n-grams corpus for finding semantically consistent distractors for these sentences. However, they do not address the challenge of choosing strategic target words, and their attempt to generate distractors at the sentence level that are contextually inconsistent at the passage level returned underwhelming results, as most target words were found to be easily distinguishable without needing previous sentences for context. While our paper addresses a similar task of finding distractors that are plausible when external context is excluded, we generate distractors at the narrower phrase level that rely on the surrounding text for context.

3 Exercise Creation

The process of automatically generating FITB exercises from an existing text involves three distinct steps: (1) choosing which target words to blank from these sentences, (2) choosing distractors for each target word, and (3) compiling these elements into a full passage-level exercise.

3.1 Choosing Target Words

The first step to creating a FITB question from a text passage is to choose which words to replace with blanks. We consider a "good" blanked question to be one in which there are enough context clues in the surrounding text for the reader to understand the text's intended meaning even when the chosen word is removed. If the reader is able to understand the sentence's intended meaning with the target word removed, then the task of replacing the word should be trivial.

We begin by considering every word in the sentence as a potential word to be blanked. However, many words would not make good target words in practice. We discard function words (articles, pronouns, conjunctions, etc.) from the pool due to their closed nature and frequent appearance across documents. However, unlike some other studies (Coniam, 1997) (Shei, 2001), we do not use global word frequencies to find uncommon words from which to create blanks. We propose that even easily-understood target words that successfully challenge comprehension of the surrounding context will implicitly test mastery of the more challenging words in the passage. However, we do consider local word frequencies, eliminating words whose stemmed form appears in the document multiple times, so that readers cannot identify target words simply by recognition due to previous encounters.

Because our exercises are designed to test understanding rather than knowledge, we also do not wish to "quiz" readers on facts, as is the case in several other studies (Karamanis et al., 2006). Therefore we also disregard classes of words that typically present factual information and could be easily exchanged for any other word of the same class (see Figure 2). These include:

- **Named entities** Specific entities, such as people, locations, and organizations
- **Numbers** Digits and their written forms

After this filtering step, the remaining set of words serves as our pool of *potential blanks*.

From this pool, we must then locate the words which relate most closely to their surrounding text. We explore several different "scopes" of context surrounding the potential blanks, and utilize word co-occurrence likelihoods to find the potential blanks that have the strongest contextual links to information within that scope. By removing words that have a meaningful contextual relationship to one or more other words in the scope, we aim to ensure that there are enough hints left remaining to enable the reader to make a reasonable inference about the blanked word.

× In fall of 2012, the New York City government began receiving unusual complaints.

× By the time California became a state, it was already an important place for farming.

Figure 2: Examples of poor target words

3.1.1 Contextual Scope

Though our goal is to generate blanks at the sentence level, individual sentences in a passage are rarely conceptually independent from one another. True understanding of a sentence's meaning often relies on information that has been gathered from previous sentences in the passage. Figure 3 gives an example of a question that relies on previous information to answer correctly.

(a) He _____ the window.
● cleaned ● opened ● saw ● closed

(b) Kahlil could not read. It was so loud outside! He _____ the window.
○ cleaned ○ opened ○ saw ● closed

Figure 3: An example of contextual scope influencing answer selection. When the word *closed* is removed (a), the reader must rely on the previous sentences to provide the context clues necessary to fill this blank (b).

Shanahan et al. (1984) found that traditional cloze-style comprehension questions are not good indicators of "intersentential comprehension," the ability to process and apply information across sentence boundaries. We therefore explore several different contextual "scopes" when attempting to find pairs of words with contextual links. Adjusting the scope of included information allows the blanks-selection method to incorporate potentially relevant or necessary context words that a reader has inter-

nalized from the sentences they have already read in order to test their intersentential comprehension.

Three scopes were tested in this paper:

s1: Context words are chosen only from the target sentence $\{s_t\}$

s2: Context words are chosen from the target sentence and the preceding sentence $\{s_{t-1}, s_t\}$

s3: Context words are chosen from the target sentence and the two preceding sentences $\{s_{t-2}, s_{t-1}, s_t\}$

The pool of scope words for each sentence is filtered less rigorously than the pool of potential blanks, as many of the word classes that make poor blanks are poor choices specifically because they provide important factual information that we wish to leverage for context. We therefore choose only to remove function words from the pool, leaving named entities, numbers, and frequently-occurring words.

3.1.2 Word Co-occurrences

We assume that words that co-occur together regularly are likely to have a contextual and/or semantic relationship to one another. We therefore utilize word co-occurrence likelihoods to select the potential blanks with the strongest relationship to their scope-specific context words.

To represent word co-occurrence likelihoods, we use the word vector space model GloVe (Pennington et al., 2014), trained on 42-billion tokens. The GloVe model formulates word vectors such that the dot product of any two word vectors $\hat{\mathbf{w_1}} \cdot \hat{\mathbf{w_2}}$ represents the logarithm of the words' probability of co-occurring together in a document.

Our goal is to find the scope word for each potential blank with the highest likelihood of co-occurring with that blanked word. Using the GloVe model, for each potential blank $b \in B$, we find the closest scope word c in the set of all scope words S for that blank such that $(b, c) = \arg\min(\hat{\mathbf{b}} \cdot \hat{\mathbf{s}})$ ($\forall s \in S$ such that $s.stem \neq b.stem$). Each of these pairs is added to the pool of blanks to carry to the task of choosing distractors.

3.2 Choosing Distractors

To turn a blanked passage into an exercise, each blank is presented as a multiple choice question. The reader is given four words to choose from that could potentially fit the given blank: the target word and three distractors. For our exercises that specifically target contextual understanding, we specify that a good distractor should make sense both grammatically and logically within a narrow context, but should not make sense within the broader context of the surrounding words.

To accomplish this, we explore a unique application of the Google Books n-grams Corpus for generating reasonable distractors for a blanked word. Google n-grams is a massive corpus containing frequency counts for all unigrams through 5-grams that occur across all texts in the Google Books corpus.

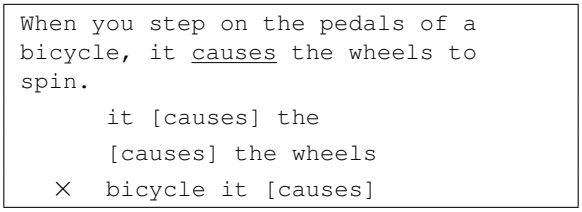

Figure 4: An example of the sentence-level trigrams extracted from a sentence. Note that an n-gram cannot occur between two clauses.

We begin by gathering every 2- through 5-gram in the original sentence that contains the target word. If the sentence contains multiple clauses, we consider only the clause which contains the target word. This allows us to avoid selecting n-grams of unusual or unintended structure (see Figure 4). We then use a sliding window to gather all n-grams ($2 \leq n \leq 5$) within the clause of the form $\{w_1...w_{t-1}, [w_t], w_{t+1}...w_n\}$, where the target word $[w_t]$ occupies each position $1 \leq t \leq n$. We then search the Google corpus for n-grams matching each pattern $\{w_1...w_{t-1}, [w_t.pos], w_{t+1}...w_n\}$ ($1 \leq t \leq n$), where $w_t.pos$ represents the part of speech of the target word w_t (obtained using the Stanford Part-Of-Speech Tagger (Toutanova et al., 2003)). If the query returns no results, we attempt to generalize the pattern further by replacing proper names and pronouns with their part of speech (see Figure 5).

We utilize a back-off model when querying for distractors, using n-grams of size $n = \{5...2\}$. For each n-sized pattern searched, we find the intersection D of all words at index t (limiting our results to the top 100 for the sake of performance).

We do not want any of the generated distractors to fit the blank as well as the target word, so we

```
James Brown [VBD] up   → ×
        NNP [VBD] up   → Moses lifted up
                         Peter stood up
                         Jill went up
                         ...
```

Figure 5: When n-gram queries return no results, we generalize specific terms to increase the likelihood of finding a match

need to remove all words in D that are likely to make too much sense in context. Because synonyms can often be used interchangeably in the same sentence, we discard all words that are direct synonyms of w_t (using synsets gathered from WordNet[1]). We also remove all words $d \in D$ such that $(\hat{\mathbf{w_t}} \cdot \hat{\mathbf{d}}) < (\hat{\mathbf{w_t}} \cdot \hat{\mathbf{c}})$ (where c is the closest scope word in the pair (w_t, c)), because these words have a *higher* likelihood than the target word does of co-occurring with their context words.

If the resulting filtered set D contains fewer than three words (the minimum required to create a multiple choice question), we back off to the next largest value of n, continuing this pattern until we have found three or more distractors for the blank. If fewer than three distractors are found after $n = 2$, the word is discarded from the pool of potential blanks. From the final set D, we select the three *least*-frequently occurring distractors in the Google corpus.

3.3 Exercise Generation

Once we have found all remaining potential blanks that have three or more distractors, we must pare down the list to create the final passage-level exercise. If any one sentence has more than one potential blank, we choose the blank that has the highest co-occurrence likelihood with its paired scope word and discard the other(s). We also discard any blanks whose paired scope word was itself made into a blank (because the context has been removed).

The resulting set of blanks constitutes the set of "best" questions for the passage. We then can present the passage-level exercise in its entirety, replacing each blanked word with a multiple choice question consisting of the target word and its three chosen distractors. Though we choose not to do so, the number of blanked sentences can also be manually limited by selecting the top x blanks from the set of all word pairs sorted by descending likelihood.

4 Evaluation

Our corpus of documents was composed of approximately 1000 reading comprehension text passages obtained from ReadWorks.org[2], ranging in reading level from 100L to 1000L using the Lexile scale. We randomly selected 120 non-unique passages (i.e. one passage could be selected more than once) from which to create questions. For each instance of a selected passage, we generated a single blanked sentence and its top three distractors, to be presented as a multiple choice question.

The questionnaire was separated into two sections, both of which asked participants to answer a set of generated FITB questions. The first section presented each question at the phrase level (i.e. the blanked surrounded by a small subset of the words in the full sentence). The words to include in these phrases were selected by hand to present the blank in a representational narrow context. The second section presented sentence-level FITB questions, surrounded by the context of the entire passage (or, in the case of particularly long passages, by relevant paragraphs from the full text). For both sections, participants were presented with four word choices for each blank, and were asked to select *all* of the words they believed logically fit the blank.

67 native English-speaking volunteers were asked to provide their feedback on each generated blank through an anonymous online questionnaire. Each participant was given a random subset of questions from each section to answer: 20 phrase-level questions, and 10 sentence-level questions. Participants were not aware that the questions were generated automatically and were not informed of the research objectives or what we hoped to obtain from their answers in order to avoid potential feedback bias.

5 Results

Alderson et al. (1995) proposed that multiple choice questions be evaluated using two metrics: *reliability* and *validity*. However, because our questions

[1]Princeton University "About WordNet." WordNet. Princeton University. 2010. <http://wordnet.princeton.edu>

[2]http://www.readworks.org/

were not answered by the target audience (i.e. low-literacy readers), we cannot compute reliability using traditional methods (such as Cronbach's alpha). We focus instead on evaluating the validity of our exercises by determining how well they conform to our proposed method of targeting narrow vs. full context.

We assess the validity of our questions and the chosen distractors by examining the proportion of words that fit each blank in a narrow context to words that fit the same blank in the broader context of the surrounding text. In an ideal question, the target word and all distractors should fit in the narrow context, and only the target word should fit given the full context. Thus, for target words, we aim for 100% fit in both contexts; for distractors, we aim for 100% fit in the narrow context and 0% in the full.

	Narrow(%)		Full(%)	
	dist	target	dist	target
$n = 2$	30.9	93.1	3.1	98.0
$n = 3$	57.7	92.4	7.0	93.8
$n = 4$	67.1	89.9	21.9	95.5
$n = 5$	74.1	93.2	13.2	91.3
ALL	58.0	92.1	11.6	94.6

Table 1: The percentage of distractors and target words chosen to fit each blank given the narrow context (left) and the full passage (right)

As can be seen in Table 1, the proportion of distractors deemed to fit the blanks in a narrow context increases substantially as n increases, while the proportion of target words chosen to fit is relatively unaffected. This pattern also holds true given the full context, although to a lesser extent.

On average, 58% of all distractors generated were deemed to fit in their given blanks in a narrow context, although this number is skewed by the poor performance of the bigram model. The 5-gram model was the best-performing for finding distractors that fit in the narrow context, achieving an average fit of approximately 74%. As n increases, more of the syntactic and semantic features of the phrase are able to be incorporated into the distractor selection, increasing the chances of the selected word making both grammatical and contextual sense with *all* of the words in the phrase.

Less than 12% of all distractors on average were deemed to fit the same blanks when given the full context, though the 4-gram model had the worst performance with nearly 22% fit. The bigram model performed best in the full context with approximately 3% fit; however, its poor performance in the narrow context suggests that these words are obviously incorrect and therefore not suitable distractors.

Table 2 compares the proportions of distractors fitting within each context across both n-gram model and scope ($s1$ through $s3$). The same pattern of increasing fit with higher values of n can be observed within each scope. However, the scope does not appear to have a significant affect on the quality of the distractors generated.

	Narrow(%)			Full(%)		
	$s1$	$s2$	$s3$	$s1$	$s2$	$s3$
$n = 2$	29.5	31.9	27.9	3.2	3.0	3.4
$n = 3$	53.7	61.2	61.0	4.6	10.2	11.1
$n = 4$	64.8	66.7	66.1	25.3	18.9	21.5
$n = 5$	75.7	74.9	75.0	13.6	13.9	20.6
ALL	56.2	59.4	56.9	10.4	11.9	13.5

Table 2: The percentage of distractors fitting each blank given the narrow (left) and full context (right), for each scope.

6 Limitations and Future Work

The proportion of words deemed to fit in the narrow contexts is lower than expected for both target words and distractors. We suspect that the concept of words "fitting" in a sentence fragment may not have been fully understood by some participants. For example, many respondents said that the word `went` was not a suitable fit for the phrase `Hidalgo _____ about this.` In this case, some participants may have struggled to identify the phrasal verb "to go about" as being grammatically correct because it clashed with the other choices (`heard`, `agreed`, `said`), where they might have chosen it to fit if it had been presented independently. A future study will explore a less subjective method of evaluating target words within a narrow context.

Perhaps the biggest weakness in our current method lies in filtering out fitting distractors. As indicated in the results above, approximately 12% of all the distractors generated using our algorithm were deemed to make as much sense in context as the target word. Upon observation, we note that the majority of the distractors chosen to fit within their full contexts are "near-synonyms" of the tar-

get word (for example, the words `turned` and `flushed`, which are not obvious synonyms but are interchangeable given the context of the phrase `her face` _____ `red`.) While we are able to remove direct synonyms using WordNet, we will work to incorporate a more robust synonym-filtering process in future work, taking advantage of the already-utilized corpora.

We also wish to further explore the relationship between scope and the target words chosen. While we have seen that adjusting the scope has little effect on the quality of the distractors generated, it remains to be seen if the target words themselves are of "better" quality for targeting comprehension as the scope of available context increases.

Alongside improvements to the question generation algorithm's performance, we also wish to prove the efficacy of these types of exercises in targeting the reading comprehension skills of low-literacy users. This process will involve further user evaluation, this time involving the target audience.

7 Conclusions

In this paper we have discussed a method of automatically generating fill-in-the-blank questions designed to target a reader's comprehension skills and contextual awareness. We have explored the idea of using word co-occurrence likelihoods coupled with scopes of context to find words with strong links to their surrounding text from which to make blanks. We have also tested a novel approach to generating distractors for these words using the Google Books n-grams corpus to find words that are semantically and logically appropriate for the given blanks in a narrow context but which do not make sense given the intention of the passage.

Results suggest that larger n-grams are significantly more effective in creating sensible distractors that make sense within a narrow context, and that a large portion of these distractors become no longer suitable once the full context of the passage has been introduced. This suggests that our method is a promising first step towards the generation of these types of comprehension-challenging exercises.

References

Manish Agarwal and Prashanth Mannem. 2011. Automatic gap-fill question generation from text books. In *Proceedings of the 6th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 56–64.

Itziar Aldabe, Montse Maritxalar, and Ruslan Mitkov. 2009. A study on the automatic selection of candidate sentences distractors. *Frontiers in Artificial Intelligence and Applications*, 200(1):656–658.

J. Charles Alderson, Caroline Clapham, and Dianne Wall. 1995. *Language Test Construction and Evaluation*.

Justin Baer, Mark Kutner, John Sabatini, and Sheida White. 2009. Basic reading skills and the literacy of america's least literate adults: Results from the 2003 national assessment of adult literacy (naal) supplemental studies. *National Center for Education Statistics*.

Jonathan C. Brown, Gwen A. Frishkoff, and Maxine Eskenazi. 2005. Automatic question generation for vocabulary assessment. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, page 826.

David Coniam. 1997. A preliminary inquiry into using corpus word frequency data in the automatic generation of english language cloze tests. *CALICO Journal*, 14(2):15–33.

Takuya Goto, Tomoko Kojiri, Toyohide Watanabe, Tomoharu Iwata, and Takeshi Yamada. 2010. Automatic generation system of multiple-choice cloze questions and its evaluation. *Knowledge Management and E-Learning*, 2(3):210–224.

Nikiforos Karamanis, Le An Ha, and Ruslan Mitkov. 2006. Generating multiple-choice test items from medical text: A pilot study. In *Proceedings of the Fourth International Natural Language Generation Conference*, number July, pages 111–113.

John Krudenier. 2002. *Research Based Principles for Adult Basic Education Reading Instruction*.

Girish Kumar, Rafael E. Banchs, and Luis F. D'Haro. 2015. Revup: Automatic gap-fill question generation from educational texts. In *Proceedings of the 10th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 154–161.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Joseph P Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A Nowak, and Erez Lieberman Aiden. 2010. Quantitative analysis of culture using millions of digitized books. *Sciencexpress*, 331(6014):176–182.

Jack Mostow and Hyeju Jang. 2012. Generating diagnostic multiple choice comprehension cloze questions.

In *Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 136–146.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543.

Juan Pino and Maxine Eskenazi. 2009. Semi-automatic generation of cloze question distractors effect of students l1. *Proceedings of the SLaTE Workshop on Speech and Language Technology in Education*, pages 1–4.

Keisuke Sakaguchi, Yuki Arase, and Mamoru Komachi. 2013. Discriminative approach to fill-in-the-blank quiz generation for language learners. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 238–242.

Timothy Shanahan, Michael L. Kamil, and Aileen Webb Tobin. 1984. Cloze as a measure of intersentential comprehension. *Reading Research Quarterly*, 14(2):229.

Chi-Chiang Shei. 2001. Followyou!: An automatic language lesson generation system. *Computer Assisted Language Learning*, 14(2):129–144.

Kristina Toutanova, Dan Klein, and Christopher D Manning. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 252–259.

Torsten Zesch and Oren Melamud. 2014. Automatic generation of challenging distractors using context-sensitive inference rules. In *Proceedings of the 9th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 143–148.