Automatic Question Generation for Vocabulary Assessment

Jonathan C. Brown
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213 USA
jonbrown@cs.cmu.edu

Gwen A. Frishkoff
Learning Research & Development Center
University of Pittsburgh
Pittsburgh, PA 15260 USA
gwenf@pitt.edu

Maxine Eskenazi
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213 USA
max@cs.cmu.edu

Abstract

In the REAP system, users are automatically provided with texts to read targeted to their individual reading levels. To find appropriate texts, the user's vocabulary knowledge must be assessed. We describe an approach to automatically generating questions for vocabulary assessment. Traditionally, these assessments have been hand-written. Using data from WordNet, we generate 6 types of vocabulary questions. They can have several forms, including wordbank and multiple-choice. We present experimental results that suggest that these automatically-generated questions give a measure of vocabulary skill that correlates well with subject performance on independently developed human-written questions. In addition, strong correlations with standardized vocabulary tests point to the validity of our approach to automatic assessment of word knowledge.

1 Introduction

The REAP system automatically provides users with individualized authentic texts to read. These texts, usually retrieved from the Web, are chosen to satisfy several criteria. First, they are selected to match the reading level of the student (Collins-Thompson and Callan, 2004). They must also have vocabulary terms known to the student. To meet this goal, it is necessary to construct an accurate model of the student's vocabulary knowledge (Brown and Eskenazi, 2004). Using this model, the system can locate documents that include a given percentage (e.g., 95%) of words that are known to the student. The remaining percentage (e.g. 5%) consists of new words that the student needs to learn. This percentage is controlled so that there is not so much stretch in the document that the student cannot focus their attention on understanding the new words and the meaning of the text. After reading the text, the student's understanding of new words is assessed. The student's responses are used to update the student model, to support retrieval of furture documents that take into account the changes in student word knowledge.

In this paper, we describe our work on automatic generation of vocabulary assessment questions. We also report results from a study that was designed to assess the validity of the generated questions. In addition to the importance of these assessments in the REAP system, tests of word knowledge are central to research on reading and language and are of practical importance for student placement and in enabling teachers to track improvements in word knowledge throughout the school year. Because tests such as these are traditionally hand-written, development is time-consuming and often relies on methods that are informal and subjective. The research described here addresses these issues through development of automated, explicit methods for generation of vocabulary tests. In addition, these tools are designed to capture the graded and complex nature of word knowledge, allowing for more fine-grained assessment of word learning.

2 Measuring Vocabulary Knowledge

Word knowledge is not all-or-none. Rather, there are different aspects, such as knowledge of the spoken form, the written form, grammatical behav-

Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), pages 819–826, Vancouver, October 2005.

ior, collocation behavior, word frequency, stylistic register constraints, conceptual meaning, and the associations a word has with other related words (Nation, 1990). In this paper, we focus on knowledge of conceptual word meaning. Because word meaning itself is complex, our focus is not simply on all-or-none estimates of vocabulary knowledge, but also on graded and incomplete knowledge of meanings that readers possess for different words and at different stages of acquisition.

Several models have been proposed to account for these multiple levels of word knowledge. For example, Dale posited four stages of knowledge of word meaning (Dale and O'Rourke, 1965). In stage 1, the subject has never seen the word. In stage 2, she has seen the word but is unable to verbalize its meaning. In stage 3, the subject recognizes the word in a given context and has partial word knowledge. In stage 4, the subject has full word knowledge, and can explain the word meaning so that its usage is clear in multiple contexts.

Stahl (1986) proposed a similar model of word knowledge, the levels of which overlap with Dale's last two stages. According to this model, the first level is characterized by association processing, or the passive association of the new word meaning with other, familiar concepts. The second level, comprehension processing, involves active comprehension of the word in a particular context. The third level, generation processing, requires usage of a word in a novel context reflecting a deep (and multidimensional) understanding of its meaning.

Taking Stahl's framework as a working model, we constructed multiple types of vocabulary questions designed to assess different "stages" or "levels" of word knowledge.

3 Question Generation

In this section, we describe the process used to generate vocabulary questions. After introducing the WordNet resource we discuss the six question types and the forms in which they appear. The use of distractors is covered in section 3.3.

3.1 WordNet

WordNet is a lexical resource in which English nouns, verbs, adjectives, and adverbs are grouped into synonym sets. A word may appear in a number of these synonym sets, or synsets, each corresponding to a single lexical concept and a single sense of the word (Fellbaum ed., 1998). The word "bat" has ten distinct senses and thus appears in ten synsets in WordNet. Five of these senses correspond to noun senses, and the other five correspond to verb senses. The synset for the verb sense of the word which refers to batting one's eyelashes contains the words "bat" and "flutter", while the synset for the noun sense of the word which refers to the flying mammal contains the words "bat" and "chiropteran". Each sense or synset is accompanied by a definition and, often, example sentences or phrases. A synset can also be linked to other synsets with various relations, including synonym, antonym, hypernym, hyponym, and other syntactic and semantic relations (Fellbaum ed., 1998). For a particular word sense, we programmatically access WordNet to find definitions, example phrases, etc.

3.2 Question Types

Given Stahl's three levels of word mastery and the information available in WordNet, we generated 6 types of questions: definition, synonym, antonym, hypernym, hyponym, and cloze questions.

In order to retrieve data from WordNet, we must choose the correct sense of the word. The system can work with input of varying specificity. The most specific case is when we have all the data: the word itself and a number indicating the sense of the word with respect to WordNet's synsets. When the target words are known beforehand and the word list is short enough, the intended sense can be hand-annotated. More often, however, the input is comprised of just the target word and its part of speech (POS). It is much easier to annotate POS than it is to annotate the sense. In addition, POS tagging can be done automatically in many cases. In the REAP system, where the user has just read a specific text, the words of the document were already automatically POS annotated. When there is only one sense of the word per part of speech, we can simply select the correct sense of the word in WordNet. Otherwise, we select the most frequently used sense of the word with the correct POS, using WordNet's frequency data. If we have only the word, we select the most frequent sense, ignoring part of speech. Future work will use word sense disambiguation techniques to automatically determine the correct word sense given a document that includes the target word, as in REAP (Brown and Eskenazi, 2004).

Once the system has determined the word sense, it can retrieve data from WordNet for each of the 6 question types. The definition question requires a definition of the word, available in WordNet's gloss for the chosen sense. The system chooses the first definition which does not include the target word. This question should provide evidence for the first of Stahl's three levels, association processing, although this was not explicitly evaluated.

The synonym question has the testee match the target word to a synonym. The system can extract this synonym from WordNet using two methods. One method is to select words that belong to the same synset as the target word and are thus synonyms. In addition, the synonym relation in WordNet may connect this synset to another synset, and all the words in the latter are acceptable synonyms. The system prefers words in the synset to those in synonym synsets. It also restricts synonyms to single words and to words which are not morphological variants of the target word. When more than one word satisfies all criteria, the most frequently used synonym is chosen, since this should make the question easier. This question could be considered either association processing or comprehension processing. If the testee has seen this synonym (e.g. as a hint), this question type would require association processing as a word is simply being associated with another already-presented word. Otherwise, this may require comprehension processing – understanding beyond memorization.

The antonym question requires matching a word with an antonymous word. WordNet provides two kinds of relations that can be used to procure antonyms: direct and indirect antonyms. Direct antonyms are antonyms of the target word, whereas indirect antonyms are direct antonyms of a synonym of the target. The words "fast" and "slow" are direct antonyms of one another. The word "quick" does not have a direct antonym, but it does have an indirect antonym, "slow", via "fast", its synonym. When more than one antonym is available, the most frequently used is chosen. Unless the testee has already seen the antonym, this type of question is normally considered to provide evidence for Stahl's second level, comprehension processing.

The hypernym and hyponym questions are similar in structure. Hypernym is the generic term used to describe a whole class of specific instances. The word "organism" is a hypernym of "person". Hyponyms are members of a class. The words "adult", "expert" and "worker" are hyponyms of "person". For the questions the testee matches the target word to either a hypernym or hyponym. For more than one possibility, the most frequently used term is chosen. Unless the testee has previously seen the hypernym or hyponym, these questions are normally regarded as providing evidence for Stahl's second level.

Cloze is the final question type. It requires the use of the target word in a specific context, either a complete sentence or a phrase. The example sentence or phrase is retrieved from the gloss for a specific word sense in WordNet. There is often more than one example phrase. The system prefers longer phrases, a feature designed to increase the probability of retrieving complete sentences. Passages using the target word are preferred, although examples for any of the words in the synset are appropriate. The present word is replaced by a blank in the cloze question phrase. Some consider a cloze question to be more difficult than any of the other question types, but it is still expected to provide evidence for Stahl's second level.

Although our question types provide evidence for the highest level of schemes such as Dale's four stages, they do not provide evidence for Stahl's highest level, generation processing, where the testee must, for instance, write a sentence using the word in a personalized context. We expect questions that provide evidence of this level to require free-form or near-free-form responses, which we do not yet allow. We expect the six question types to be of increasing difficulty, with definition or synonym being the easiest and cloze the hardest.

3.3 Question Forms

Each of the 6 types of questions can be generated in several forms, the primary ones being wordbank and multiple-choice. In wordbank, the testee sees a list of answer choices, followed by a set of questions or statements (see Figure 1). For the definition version, each of the items below the wordbank is a definition. The testee must select the word which best corresponds to the definition. For the synonym and antonym questions, the testee selects the word which is the most similar or the most opposite in meaning to the synonym or antonym. For the hypernym and hyponym question types, the testee is asked to complete phrases such as "___ is a kind of person" (with target "adult") or "person

is a kind of ___" (with target "organism"). In the cloze question, the testee fills in the blank with the appropriate word. There is traditionally one question for each target word in the wordbank. These questions require no information beyond the target words and their definitions, synonyms, hypernyms, etc.

Wordbank:

verbose infallible obdurate opaque

Choose the word from the wordbank that best completes each phrase below:

1. ___ windows of the jail
2. the Catholic Church considers the Pope ___
3. ___ and ineffective instructional methods
4. the child's misery would move even the most ___ heart

Fig. 1. Example Wordbank Question

The second generated form is multiple-choice, with one question per target word. The testee sees the main question, the stem, followed by several answer choices, of which only one is correct (see Figure 2). Depending on the question type, the target word may appear in either the stem or the answer choices. For the definition question type, the stem holds the definition of the target word and one of the answer choices is the target word. For the word "verbose", the stem would be "using or containing too many words" and the choices "ancillary", "churlish", "verbose", and "convivial". The cloze question is of a similar form, with the stem containing the example sentence or phrase with a blank where the target word should be used. For "verbose", we have the stem "___ and ineffective instructional methods" and choices "verbose", "incipient", "invidious", and "titular". For the synonym, antonym, hypernym, and hyponym questions, the target word appears in the stem instead of the answer choices. The synonym question for the word "verbose" would have the stem "Select the word that is most similar in meaning to the word verbose" with choices "inflammable", "piping", matrilineal", and "long-winded". The antonym question would have the stem "Select the word that is most opposite in meaning to the word verbose" and the choices "discernable", "concise", "unbroken", and "soused". Figure 2 shows a formatted example of an automatically generated multiple-choice cloze question for the word "obdurate".

Choose the word that best completes the phrase below:

the child's misery would move even the most ___ heart

A) torpid
B) invidious
C) stolid
D) obdurate

Fig. 2. Example Multiple-Choice Cloze Question

Two issues to consider when creating multiple-choice format questions are the wording or appearance of the questions and the criteria for selection of distractors. We followed the guidelines for good multiple-choice questions described by researchers such as Graesser and Wisher (2001). In accord with these guidelines, our questions had 4 choices, although the number of choices is a variable supplied to the question generation software. We also considered the most appropriate wording for these questions, leading us to choose stems such as "Select the word that is most similar in meaning to the word *plausible*" for the synonym question rather than "Choose the word that means the same as the word *plausible*." The latter would be problematic when the correct answer is a near-synonym rather than a word with precisely the same meaning.

Concerning distractor choice, the question generation system chooses distractors of the same part of speech and similar frequency to the correct answer, as recommended by Coniam (1997). For the synonym, antonym, hypernym, and hyponym questions, the correct answer is the highest frequency word of *all* the words chosen from WordNet that satisfy all the criteria. Thus, the distractors are of the same POS and similar frequency to the synonym, antonym, or whatever word is the correct answer, as opposed to the target word. The system chooses distractors from Kilgarriff's (1995) word frequency database, based on the British National Corpus (BNC) (Burnage, 1991). The system chooses 20 words from this database that are of the same POS and are equal or similar in frequency to the correct answer, and randomly chooses the distractors from these words. Since the distractors may be different for each run of the question generation software, slightly different versions of the same basic question may appear. The words of the BNC and the word frequency database have been POS tagged using the CLAWS tagger (Leech, 1994). This tagger uses detailed POS tags, enabling us to choose distractors that are, for instance,

verbs in the past tense, when the correct answer is such as verb, instead of selecting verbs of unknown tense. In the definition and cloze questions, the correct answer is the target word itself, so distractors are chosen based on this word. The system also restricts distractors to be in the list of target words so that the testee cannot simply choose the word that appears in the stems of other questions.

An alternate multiple-choice question format is used when the testee has just read a document using the target word, as in the REAP system (Brown and Eskenazi, 2004). In this case, the system also attempts to finds words which may be semantically related to the correct answer, as in (Nagy, 1985). This is done by choosing distractors that satisfy the standard criteria and were present in the document. This should increase the chance that the distractors are semantically related and eliminate the chance that a testee will simply select as the correct answer the word that appeared in the document they just read, without understanding the word meaning.

4 Question Assessment

The validity of the automatically generated vocabulary questions was examined in reference to human-generated questions for 75 low-frequency English words. We compared student performance (accuracy and response time) on the computer and human-generated questions. We focused on the automatically generated multiple-choice questions, with distractors based on frequency and POS. We did not examine using more complicated strategies for picking distractors or assume there was an associated text. Four of the six computer-generated question types were assessed: the definition, synonym, antonym, and cloze questions. Hypernym and hyponym questions were excluded, since we were unable to generate a large number of these questions for adjectives, which constitute a large portion of the word list. Subject scores on the computer and human-generated assessments were compared with scores on standardized measures of reading and vocabulary skill, as described below.

4.1 Question Coverage

Potential experimental stimuli comprised 156 low-frequency and rare English words that have been used in previous studies of vocabulary skill in native English-speaking adults. We first examined the percentage of words for which we could generate various question types. We were unable to generate any questions for 16 of these words, or ~9% of the list, since they were not in WordNet. Table 1 shows the percentage of words for which each of the four question types was generated. All four questions were able to be generated for only 75 (about half) of the words. Therefore, the experimental word list included only these 75 items. Given the rarity of the words, we predicted that the percentage of words for which we could generate questions would be lower than average. However, we expected that the percentage of words for which we could generate synonym and antonym questions to be higher than average, due to the heavy focus on adjectives in this list.

Question type	Percentage of Questions Generated
Definition Question	91%
Synonym Question	80%
Antonym Question	60%
Cloze Question	60%

Table 1. Question Coverage for the 156-Word List

4.2 Experiment Design

Behavioral measures of vocabulary knowledge were acquired for the 75 target words using the four computer-generated question types described above, as well as five human-generated question types. The human-generated questions were developed by a group of three learning researchers, without knowledge of the computer-generated question types. Researchers were asked merely to develop a set of question types that could be used to assess different levels, or different aspects, of word knowledge. Examples of each question type (including distractors) were hand-written for each of the 75 words.

Two of the five human-generated assessments, the synonym and cloze questions, were similar in form to the corresponding computer-generated question types in that they had the same type of stem and answer. The other three human-generated questions included an inference task, a sentence completion task, and a question based on the Osgood semantic differential (Osgood, 1970). In the inference task, participants were asked to select a context where the target word could be meaningfully applied. For example, the correct response to

the question "Which of the following is most likely to be lenitive?" was "a glass of iced tea," and distractors were "a shot of tequila," "a bowl of rice," and "a cup of chowder." In the sentence completion task, the participant was presented with a sentence fragment containing the target word and was asked to choose the most probable completion. For example, the stem could be "The music was so lenitive…," with the correct answer "…it was tempting to lie back and go to sleep," and with distractors such as "…it took some concentration to appreciate the complexity." The fifth question type was based on the Osgood semantic differential, a factor-analytic model of word-level semantic dimensions (Osgood, 1970). Numerous studies using the Osgood paradigm have shown that variability in the semantic "structure" of word meanings can largely be accounted for in terms of three dimensions, valence (good–bad), potency (strong–weak), and activity (active–passive). In our version of the Osgood task, subjects were asked to classify a word such as "lenitive" along one of these dimensions (e.g., more good or more bad).

In addition to the human-generated questions, we administered a battery of standardized tests, including the Nelson-Denny Reading Test, the Raven's Matrices Test, and the Lexical Knowledge Battery. The Nelson-Denny Reading Test is a standardized test of vocabulary and reading comprehension (Brown, 1981). The Raven's Matrices Test is a test of non-verbal reasoning (Raven, 1960). The Lexical Knowledge Battery has multiple subsections that test orthographic and phonological skills (Perfetti and Hart, 2001).

Twenty-one native-English speaking adults participated in two experiment sessions. Session 1 lasted for about one hour and included the battery of vocabulary and reading-related assessments described above. Session 2 lasted between two and three hours and comprised 10 tasks, including the five human and four computer-generated questions. The experiment began with a confidence-rating task, in which participants indicated with a key press how well they knew the meaning of each target word (on a 1–5 scale). This task was not speeded. For the remaining tasks, subjects were asked to respond "as quickly as possible without making errors." Test items for a given question type were answered together. The order of the tasks (question types) and the order of the 75 items within each task were randomized across subjects.

4.3 Experiment Results

We report on four aspects of this study: participant performance on questions, correlations between question types, correlations with confidence ratings, and correlations with external assessments.

Mean accuracy scores for each question type varied from .5286 to .6452. Performance on individual words and across subjects (averaging across words) varied widely. The easiest question types (those with the highest average accuracy), were the computer-generated definition task and the human-generated semantic differential task, both having mean accuracy scores of .6452. The hardest was the computer-generated cloze task, with a mean score of .5286. The accuracy on computer-generated synonym and antonym questions fall between these two limits, with slightly greater accuracy on the synonym type. This implies a general ordering of difficulty from definition to cloze, as expected. The accuracies on the other human-generated questions also fall into this range.

We also computed correlations between the different question types. Mean accuracies were highly and statistically significantly correlated across the nine question types ($r>.7$, $p<.01$ for all correlations). The correlation between participant accuracy on the computer-generated synonym and the human-generated synonym questions was particularly high ($r=.906$), as was the correlation between the human and computer cloze questions ($r= .860$). The pattern of correlations for the response-time (RT) data was more complicated and is discussed elsewhere (Frishkoff et al, In Prep). Importantly, RTs for the human versus computer versions of both the synonym and cloze questions were strongly correlated ($r>.7$, $p<.01$), just as for the accuracy results. The accuracy correlations imply that the computer-generated questions are giving a measure of vocabulary skill for specific words that correlates well with that of the human-generated questions.

An item analysis (test item discrimination) was also performed. For each word, scores on a particular question type were compared with the composite test score for that word. This analysis revealed relatively low correlations ($.12 < r < .25$) between the individual question types and the test as a whole (without that question type). Since the question types were designed to test different aspects of vocabulary knowledge, this result is encouraging.

In addition, the average total-score correlations for the four computer-generated questions (r=.18) and for the five human-generated questions (r=.19) were not significantly different. This is positive, since it suggests that the human and computer-generated vocabulary test are accounting for similar patterns of variance across the different question types.

The average correlation between accuracy on the question types and confidence ratings for a particular word was .265. This correlation was unexpectedly low. This may be because participants thought they knew these words, but were confused by their rarity, or because confidence simply does not correlate well with accuracy. Further work is needed to determine whether confidence ratings can be accurate predictors of vocabulary knowledge.

Finally, we examined correlations between participant performance on the nine question types and the external assessments. The correlations between the accuracy on each of the nine question types and the Nelson-Denny vocabulary subtest were fairly high ($.61 < r < .85$, $p=.01$ for all comparisons). Thus, both the computer and human-generated questions show good correspondence with an external assessment of vocabulary skill. Correlations between the accuracy on the question types and the Nelson-Denny reading comprehension test were mixed, showing a higher correlation with vocabulary than reading comprehension. Correlations between the accuracy on the nine question types and the Raven's Matrices test of nonverbal reasoning were positive, but low and not statistically significant. This provides strong evidence that the computer-generated vocabulary questions tap vocabulary knowledge specifically, rather than intelligence in general.

5 Related Work

Cloze tests are one area of related work. They were originally intended to measure text readability (Taylor, 1953) since native speakers should be able to reproduce certain removed words in a readable text. Other researchers have used it to assess reading comprehension (Ruddell, 1964), with students filling in the blanks, given a high quality text. The main issue in automating the creation of cloze tests is determining which words to remove from the text. Coniam (1997) examined a several options for determining the words to remove and produced relatively good-quality cloze tests by removing words with the same POS or similar frequency.

Wolfe (1976) automatically generated reading comprehension questions. This involved various techniques for rewriting sentences into questions, testing syntactic understanding of individual sentences. Of the 50 questions Wolfe was able to generate for a single text, 34 were found to be satisfactory. More recently, Kunichika (2003) carried out work in automatically generating reading comprehension questions that included both syntactic and semantic questions, and was able to generate several different types of questions, including asking about the content of a sentence, using dictionaries of synonyms and antonyms to generate questions such as "Is Jane busy?" from sentences like "Jane is free.", and testing semantic understanding across sentence boundaries. Approx. 93% of the generated questions were found to be satisfactory.

Aist (2001) automatically generated factoids to assist students reading. The factoids gave a synonym, an antonym, or a hypernym for the word, which were automatically extracted from WordNet. He also automated the creation of a single type of vocabulary question, with the target word in the stem and the correct answer a synonym, hypernym, or sibling from WordNet. It is unclear what type of vocabulary knowledge this question would tap, given the different possible answers.

6 Conclusions

Extending our experiments to the question types that we have not yet assessed is an important next step. In addition, we want to assess questions individually, evaluating their use of distractors. Finally, we need to assess questions generated on word lists with different characteristics.

There are also a number of ongoing extensions to this project. One is the creation of new question types to test other aspects of word knowledge. Another is using other resources such as text collections to enable us to generate more questions per word, especially for the cloze questions. In addition, we are looking at ways to predict word knowledge using confidence ratings and morphological and semantic cohorts in situations where we cannot perform a standard assessment or cannot test all the vocabulary words we would like to.

In this paper, we have described our work in automatically generating questions for vocabulary assessment. We have described the six types of computer-generated questions and the forms in which they appear. Finally, we have presented evidence that the computer-generated questions give a measure of vocabulary skill for individual words that correlates well with human-written questions and standardized assessments of vocabulary skill.

Acknowledgements

The authors would like to thank Jamie Callan and Kevyn Collins-Thompson for their help in this research. The authors would also like to thank Eve Landen, Erika Taylor, and Charles Perfetti for their assistance with experimental stimuli and data collection. This project is supported U.S. Department of Education, Institute of Education Sciences, Award #R305G030123, and the APA/IES Postdoctoral Education Research Training fellowship awarded by the Department of Education, Institute of Education Sciences, Grant #R305U030004. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the U.S. Department of Education.

References

Gregory Aist. 2001. Towards automatic glossarization: automatically constructing and administering vocabulary assistance factoids and multiple-choice assessment, *International Journal of AI in Ed.*, 2001.

James Brown, J. M. Bennett, and Gerald Hanna. 1981. *The Nelson-Denny Reading Test*. Chicago: The Riverside Publishing Company.

Jonathan Brown and Maxine Eskenazi. 2004. Retrieval of Authentic Documents for Reader-Specific Lexical Practice. In *Proceedings of InSTIL/ICALL Symposium 2004*. Venice, Italy, 2004.

Gavin Burnage. 1991. Text Corpora and the British National Corpus. *Computers & Texts* 2, Nov, 1991.

Kevyn Collins-Thompson and Jamie Callan. 2004. A language modeling approach to predicting reading difficulty. In *Proceedings of the HLT/NAACL 2004 Conference*. Boston, 2004.

David Coniam. 1997. A preliminary inquiry into using corpus word frequency data in the automatic generation of English language cloze tests. *CALICO Journal*, Volume 14, No. 2.

Edgar Dale and Joseph O'Rourke. 1986. *Vocabulary building*. Columbus, Ohio: Zaner-Bloser.

Christiane Fellbaum, Ed. 1998. *WordNet. An electronic lexical database*. Ed. by Christiane Fellbaum, preface by George Miller. Cambridge, MA: MIT Press; 1998.

Arthur C. Graesser, R. A. Wisher. 2001. *Question Generation as a Learning Multiplier in Distributed Learning Environments*. Army research inst for the behavioral and social sciences Alexandria VA. Report number A654993, 2001.

Adam Kilgarriff. 1995. http://www.itri.brighton.ac.uk/~Adam.Kilgarriff/bnc-readme.html

H. Kunichika, T. Katayama, T. Hirashima, and A. Takeuchi. 2003. Automated question generation methods for intelligent English learning systems and its evaluation. *Proceedings of ICCE2004*, Hong Kong.

G. Leech, R. Garside, and M. Bryant. 1994. CLAWS4: The tagging of the British National Corpus. In *Proc. of 15th International Conference on Computational Linguistics*, Kyoto, Japan, 622-628, 1994.

W.E. Nagy, P.A. Herman, and R.C. Anderson. 1985. Learning words from context. *Reading Research Quarterly*, 20, 233-253.

Paul Nation. 1990. *Teaching and learning vocabulary*. Rowley, MA: Newbury House.

Charles E. Osgood, P. H. Tannenbaum, and G. J. Suci. 1957. *The Measurement of Meaning*. Urbana: University of Illinois Press.

Charles A. Perfetti, and Lesley Hart. 2001. The lexical quality hypothesis. In L. Verhoeven, C. Elbro & P. Reitsma (Eds.), *Precursors of Functional Literacy* (Vol. 11, pp. 67–86). Amsterdam: John Benjamins.

J.C. Raven. 1960. *Progressive matrices, standard*. San Antonio, TX: Psychological Corporation.

R. B. Ruddell. 1964. A study of the cloze comprehension technique in relation to structurally controlled reading material. *Improvement of Reading Through Classroom Practice*, 9, 298-303.

Steven A. Stahl. 1986. Three principals of effective vocabulary instruction. *Journal of Reading*, 29.

W.L. Taylor. 1953. Cloze procedure: a new tool for measuring readability. *Journalism Quarterly*, 30.

John H. Wolfe. 1976. Automatic question generation from text - an aid to independent study. *ACM SIGCUE Bulletin*, 2(1), 104-112.