Predicting Cloze Task Quality for Vocabulary Training

Adam Skory **Maxine Eskenazi**
Language Technologies Institute
Carnegie Mellon University
Pittsburgh PA 15213, USA
{askory,max}@cs.cmu.edu

Abstract

Computer generation of cloze tasks still falls short of full automation; most current systems are used by teachers as authoring aids. Improved methods to estimate cloze quality are needed for full automation. We investigated lexical reading difficulty as a novel automatic estimator of cloze quality, to which co-occurrence frequency of words was compared as an alternate estimator. Rather than relying on expert evaluation of cloze quality, we submitted open cloze tasks to workers on Amazon Mechanical Turk (AMT) and discuss ways to measure of the results of these tasks. Results show one statistically significant correlation between the above measures and estimators, which was lexical co-occurrence and Cloze Easiness. Reading difficulty was not found to correlate significantly. We gave subsets of cloze sentences to an English teacher as a gold standard. Sentences selected by co-occurrence and Cloze Easiness were ranked most highly, corroborating the evidence from AMT.

1 Cloze Tasks

Cloze tasks, described in Taylor (1953), are activities in which one or several words are removed from a sentence and a student is asked to fill in the missing content. That sentence can be referred to as the 'stem', and the removed term itself as the 'key'. (Higgins, 2006) The portion of the sentence from which the key has been removed is the 'blank'. 'Open cloze' tasks are those in which the student can propose any answer. 'Closed cloze' describes multiple choice tasks in which the key is presented along with a set of several 'distractors'.

1.1 Cloze Tasks in Assessment

Assessment is the best known application of cloze tasks. As described in (Alderson, 1979), the "cloze procedure" is that in which multiple words are removed at intervals from a text. This is mostly used in first language (L1) education. Alderson describes three deletion strategies: random deletion, deletion of every n^{th} word, and targeted deletion, in which certain words are manually chosen and deleted by an instructor. Theories of lexical quality (Perfetti & Hart, 2001) and word knowledge levels (Dale, 1965) illustrate why cloze tasks can effectively assess multiple dimensions of vocabulary knowledge.

Perfetti & Hart explain that lexical knowledge can be decomposed into orthographic, phonetic, syntactic, and semantic constituents. The lexical quality of a given word can then be defined as a measure based on both the depth of knowledge of each constituent and the degree to which those constituents are bonded together. Cloze tasks allow a test author to select for specific combinations of constituents to assess (Bachman, 1982).

1.2 Instructional Cloze Tasks

Cloze tasks can be employed for instruction as well as assessment. Jongsma (1980) showed that targeted deletion is an effective use of instructional passage-based cloze tasks. Repeated exposure to frequent words leads first to familiarity with those words, and increasingly to suppositions about their semantic and syntactic constituents. Producing cloze tasks through targeted deletion takes implicit, receptive word knowledge, and forces the student

to consider explicitly how to match features of the stem with what is known about features of any keys she may consider.

2 Automatic Generation of Cloze Tasks

Most cloze task "generation" systems are really cloze task *identification* systems. That is, given a set of requirements, such as a specific key and syntactic structure (Higgins 2006) for the stem, a system looks into a database of pre-processed text and attempts to identify sentences matching those criteria. Thus, the content generated for a closed cloze is the stem (by deletion of the key), and a set of distractors. In the case of some systems, a human content author may manually tailor the resulting stems to meet further needs.

Identifying suitable sentences from natural language corpora is desirable because the sentences that are found will be authentic. Depending on the choice of corpora, sentences should also be well-formed and suitable in terms of reading level and content. Newspaper text is one popular source (Hoshino & Nakagawa, 2005; Liu et al., 2005; Lee & Seneff, 2007). Pino et al. (2008) use documents from a corpus of texts retrieved from the internet and subsequently filtered according to readability level, category, and appropriateness of content. Using a broader corpus increases the number and variability of potential matching sentences, but also lowers the confidence that sentences will be well-formed and contain appropriate language (Brown & Eskenazi, 2004).

2.1 Tag-based Sentence Search

Several cloze item authoring tools (Liu et al. 2005; Higgins, 2006) implement specialized tag-based sentence search. This goes back to the original distribution of the Penn Treebank and the corresponding *tgrep* program. Developed by Pito in 1992 (Pito, 1994) this program allows researchers to search for corpus text according to sequences of part of speech (POS) tags and tree structure.

The linguists' Search Engine (Resnik & Elkiss, 2005) takes the capabilities of *tgrep* yet further, providing a simplified interface for linguists to search within tagged corpora along both syntactic and lexical features.

Both *tgrep* and the Linguists' Search Engine were not designed as cloze sentence search tools, but they paved the way for similar tools specialized for this task. For example, Higgins' (2006) system uses a regular expression engine that can work either on the tag level, the text level or both. This allows test content creators to quickly find sentences within very narrow criteria. They can then alter these sentences as necessary.

Liu et al. (2005) use sentences from a corpus of newspaper text tagged for POS and lemma. Candidate sentences are found by searching on the key and its POS as well as the POS sequence of surrounding terms. In their system results are filtered for proper word sense by comparing other words in the stem with data from WordNet and HowNet, databases of inter-word semantic relations.

2.2 Statistical Sentence Search

Pino et al (2009) use co-occurrence frequencies to identify candidate sentences. They used the Stanford Parser (Klein & Manning, 2003) to detect sentences within a desired range of complexity and likely well-formedness. Co-occurrence frequencies of words in the corpus were calculated and keys were compared to other words in the stem to determine cloze quality, producing suitable cloze questions 66.53% of the time. This method operates on the theory that the quality of the context of a stem is based on the co-occurrence scores of other words in the sentence. Along with this result, Pino et al. incorporated syntactic complexity in terms of the number of parses found.

Hoshino & Nakagawa (2005) use machine learning techniques to train a cloze task search system. Their system, rather than finding sentences suitable for cloze tasks, attempts to automate deletion for passage-based cloze. The features used include sentence length and POS of keys and surrounding words. Both a Naïve Bayes and a K-Nearest Neighbor classifier were trained to find the most likely words for deletion within news articles. To train the system they labeled cloze sentences from a TOEIC training test as *true*, then shifted the position of the blanks from those sentences and

labeled the resulting sentences as *false*. Manual evaluation of the results showed that, for both classifiers, experts saw over 90% of the deletions as either easy to solve or merely possible to solve.

3 Reading Level and Information Theory

An information-theoretical basis for an entirely novel approach to automated cloze sentence search is found in Finn (1978). Finn defines *Cloze Easiness* as "the percent of subjects filling in the correct word in a cloze task." Another metric of the quality of a cloze task is context restriction; the number of solutions perceived as acceptable keys for a given stem. Finn's theory of lexical feature transfer provides one mechanism to explain context restriction. The theory involves the *information* content of a blank.

According to Shannon's (1948) seminal work on information theory, the *information* contained in a given term is inverse to its predictability. In other words, if a term appears despite following a history after which is it considered very unlikely to occur, that word has high *information* content. For example, consider the partial sentence *"She drives a nice..."*. A reader forms hypotheses about the next word before seeing it, and thus expects an overall meaning of the sentence. A word that conforms to this hypothesis, such as the word 'car', does little to change a reader's knowledge and thus has little *information*. If instead the next word is 'taxi', 'tank', or 'ambulance', unforeseen knowledge is gained and relative *information* is higher.

According to Finn (1978) the applicability of this theory to Cloze Easiness can be explained though lexical transfer features. These features can be both syntactic and semantic, and they serve to interrelate words within a sentence. If a large number of lexical transfer features are within a given proximity of a blank, then the set of words matching those features will be highly restricted. Given that each choice of answer will be from a smaller pool of options, the probability of that answer will be much higher. Thus, a highly probable key has correspondingly low *information* content.

Predicting context restriction is of benefit to automatic generation of cloze tasks. Cloze Easiness improves if a student chooses from a smaller set of possibilities. The instructional value of a highly context-restricted cloze task is also higher by providing a richer set of lexical transfer features with which to associate vocabulary.

Finn's application of information theory to Cloze Easiness and context restriction provides one possible new avenue to improve the quality of generated cloze tasks. We hypothesize that words of higher reading levels contain higher numbers of transfer features and thus their presence in a sentence can be correlated with its degree of context restriction. To the authors' knowledge reading level has not been previously applied to this problem.

We can use a unigram reading level model to investigate this hypothesis. Returning to the example words for the partial sentence *"She drives a nice..."*, we can see that our current model classifies the highly expected word, 'car', at reading level *1*, while 'taxi','tank', and 'ambulance', are at reading levels *5*, *6*, and *11* respectively.

3.1 Reading Level Estimators

The estimation of reading level is a complex topic unto itself. Early work used heuristics based on average sentence length and the percentage of words deemed unknown to a baseline reader. (Dale & Chall, 1948; Dale, 1965) Another early measure, the Flesch-Kincaid measure, (Kincaid et al., 1975) uses a function of the syllable length of words in a document and the average sentence length.

More recent work on the topic also focuses on readability classification at the document level. Collins-Thompson & Callan (2005) use unigram language models without syntactic features. Heilman et al. (2008) use a probabilistic parser and unigram language models to combine grammatical and lexical features. (Petersen & Ostendorf, 2006) add higher-order n-gram features to the above to train support vector machine classifiers for each grade level.

These recent methods perform well to characterize the level of an entire document, but they are untested for single sentences. We wish to investigate if a robust unigram model of reading level can be employed to improve the estimation of cloze quality at the sentence level. By extension of Finn's (1978) hypothesis, it is in fact not the

overall level of the sentence that has a predicted effect on cloze context restriction, but rather the reading level of the words in proximity to the blank. Thus we propose that it should be possible to find a correlation between cloze quality and the reading levels of words in near context to the blank of a cloze task.

4 The Approach

We investigate a multi-staged filtering approach to cloze sentence generation. Several variations of the final filtering step of this approach were employed and correlations sought between the resulting sets of each filter variation. The subset predicted to contain the best sentences by each filter was finally submitted to expert review as a gold standard test of cloze quality.

This study compares two features of sentences, finding the levels of context restriction experimentally. The first feature in question is the maximum reading level found in near-context to the blank. The second feature is the mean skip bigram co-occurrence score of words within that context.

Amazon Mechanical Turk (AMT) is used as a novel cloze quality evaluation method. This method is validated by both positive correlation with the known-valid (Pino et al., 2008) co-occurrence score predictor, and an expert gold standard. Experimental results from AMT are then used to evaluate the hypothesis that reading level can be used as a new, alternative predictor of cloze quality.

4.1 Cloze Sentence Filtering

The first step in preparing material for this study was to obtain a set of keys. We expect that in most applications of sentence-based cloze tasks the set of keys is pre-determined by instructional goals. Due to this constraint, we choose a set of keys distributed across several reading levels and hold it as fixed. Four words were picked from the set of words common in texts labeled as grades four, six, eight, ten, and twelve respectively.

4th: 'little', 'thought', 'voice', 'animals'
6th: ‘president', 'sportsmanship', 'national', experience'
8th: 'college', 'wildlife', 'beautiful', 'competition'
10th: 'medical', 'elevations','qualities', 'independent'
12th: 'scientists', 'citizens', 'discovered', 'university'

Figure 1: common words per grade level.

201,025 sentences containing these keys were automatically extracted from a corpus of web documents as the initial filtering step. This collection of sentences was then limited to sentences of length 25 words or less. Filtering by sentence length reduced the set to 136,837 sentences.

A probabilistic parser was used to score each sentence. This parser gives log-probability values corresponding to confidence of the best parse. A threshold for this confidence score was chosen manually and sentences with scores below the threshold were removed, reducing the number of sentences to 29,439.

4.2 Grade Level

Grade level in this study is determined by a smoothed unigram model based on normalized concentrations within labeled documents. A sentence is assigned the grade level of the highest level word in context of the key.

4.3 Co-occurrence Scores

Skip bigram co-occurrence counts were calculated from the Brown (Francis & Kucera, 1979) and OANC (OANC, 2009) corpora. A given sentence's score is calculated as the mean of the probabilities of finding that sentence's context for the key.

These probabilities are defined on the triplet (*key*, *word*, *window size*), in which *key* is the target word to be removed, *word* any term in the corpus, and *window size* is a positive integer less than or equal to the length of the sentence.

This probability is estimated as the number of times *word* is found within the same sentence as *key* and within an absolute *window size* of *2* positions from *key*, divided by the total number of times all terms are found in that window. These scores are thus maximum likelihood estimators of the probability of *word* given *key* and *window size:*

(1) For some key *k* , word *w*, and window-size *m* :

$C_j(w, k)$:= count of times *w* found *j* words from the position of *k*, within the same sentence.

(2) For a vocabulary *V* and for some positive integer window-size *m*, let *n = (m-1) / 2*, then:

$$P(w|k,m) = \frac{\sum_{j\in[-n,n],j\neq 0} C_j(w,k)}{\sum_{t_i\in V}\sum_{j\in[-n,n],j\neq 0} C_j(t_i,k)}$$

i.e. if our corpus consisted of the single sentence *"This is a good example sentence."*:

C−1 (w = good, *k* = example*) = 1*
C1 (w = sentence, *k* = example*) = 1*
P (w = good | *k* = example, *m = 3) = 1 / (1+1)= .5*

Finally, the overall score of the sentence is taken to be the mean of the skip bigram probabilities of all words in context of the key.

4.4 Variable Filtering by Grade and Score

Skip bigram scores were calculated for all words co-occurrent in a sentence with each of our 20 keys. To maximize the observable effect of the two dimensions of grade level and co-occurrence score, the goal was to find sentences representing combinations of ranges within those dimensions. To achieve this it was necessary to pick the window size that best balances variance of these dimensions with a reasonably flat distribution of sentences.

In terms of grade level, smaller window sizes resulted in very few sentences with at least one high-level word, while larger window sizes resulted in few sentences with no high-level words. Variance in co-occurrence score, on the other hand, was maximal at a window size of 3 words, and dropped off until nearly flattening out at a window size of 20 words. A window size of 15 words was found to offer a reasonable distribution of grade level while preserving sufficient variance of co-occurrence score.

Using the above window-size, we created filters according to maximum grade level: one each for the grade ranges 5-6, 7-8, 9-10, and 11-12. Four more filters were created according to co-occurrence score: one selecting the highest-scoring quartile of sentences, one the second highest-scoring quartile, and so on. Each grade level filter was combined with each co-occurrence score filter creating 4x4=16 composite filters. By combining these filters we can create a final set of sentences for analysis with high confidence of having a significant number of sentences representing all possible values of grade level and co-occurrence score. At most two sentences were chosen for each of the 20 keys using these composite filters. The final number of sentences was 540.

4.5 Experimental Cloze Quality

Previous evaluation of automatically generated cloze tasks has relied on expert judgments. (Pino et al., 2008; Liu et al., 2005) We present the use of crowdsourcing techniques as a new approach for this evaluation. We believe the approach can be validated by statistically significant correlations with predicted cloze quality and comparison with expert judgments.

The set of 540 sentences were presented to workers from Amazon Mechanical Turk (AMT), an online marketplace for "human intelligence tasks." Each worker was shown up to twenty of the stems of these sentences as open cloze tasks. No worker was allowed to see more than one stem for the same key. Workers were instructed to enter only those words that "absolutely make sense in this context", but were not encouraged to submit any particular number of answers. Workers were paid US$.04 per sentence, and the task was limited to workers with approval ratings on past tasks at or above 90%.

For each sentence under review each worker contributes one subset of answers. Cloze Easiness, as defined by Finn (1978) is calculated as the percentage of these subsets containing the original key. We define *context restriction* on *n* as the percentage of answer subsets containing *n* or fewer words.

Using the example sentence: "Take this cloze sentence, for __(example)__." We can find the set of answer subsets A:

$$A = \{\ A_1=\{\text{example, free, fun, me}\}$$
$$A_2=\{\text{example,instance}\}$$
$$A_3=\{\text{instance}\}\ \}$$

Then, Cloze Easiness is $|\{A_1,A_2\}| / |A| \approx .67$ and Context restriction (on one or two words) is $|\{A_2,A_3\}| / |A| \approx .67$

5 Results

Each sentence in the final set was seen, on average, by 27 Mechanical Turk workers. We wish to correlate measures of Cloze Easiness and context restriction with cloze quality predictors of maximum grade level and score. We use the Pearson correlation coefficient (PCC) to test the linear relationship between each measure of cloze quality and each predictor.

Table (1) shows these PCC values. All of the values are positive, meaning there is a correlation showing that one value will tend to increase as the other increases. The strongest correlation is that of co-occurrence and Cloze Easiness. This is also the only statistically significant correlation. The value of $P(H_0)$ represents the likelihood of the null hypothesis: that two random distributions generated the same correlation. Values of $P(H_0)$ under 0.05 can be considered statistically significant.

Cloze Easiness	PCC = 0.2043 $P(H_0)$=1.6965e-06	PCC = 0.0671 $P(H_0)$=0.1193
Context Restriction (2)	PCC = 0.0649 $P(H_0)$=0.1317	PCC = 0.07 $P(H_0)$=0.1038
	Co-occurrence	Maximum Grade

Table (1): Pearson Correlation Coefficient and probability of null hypothesis for estimators and measures of cloze quality.

Figure (3) shows scatter plots of these four correlations in which each dot represents one sentence.

The top-leftmost plot shows the correlation of co-occurrence score (on the x-axis), and Cloze Easiness (on the y-axis). Co-occurrence scores are shown on a log-scale. The line through these points represents a linear regression, which is in this case statistically significant.

The bottom-left plot shows correlation of co-occurrence score (x-axis) with context restriction. In this case context restriction was calculated on *n=2*, i.e. the percent of answers containing only

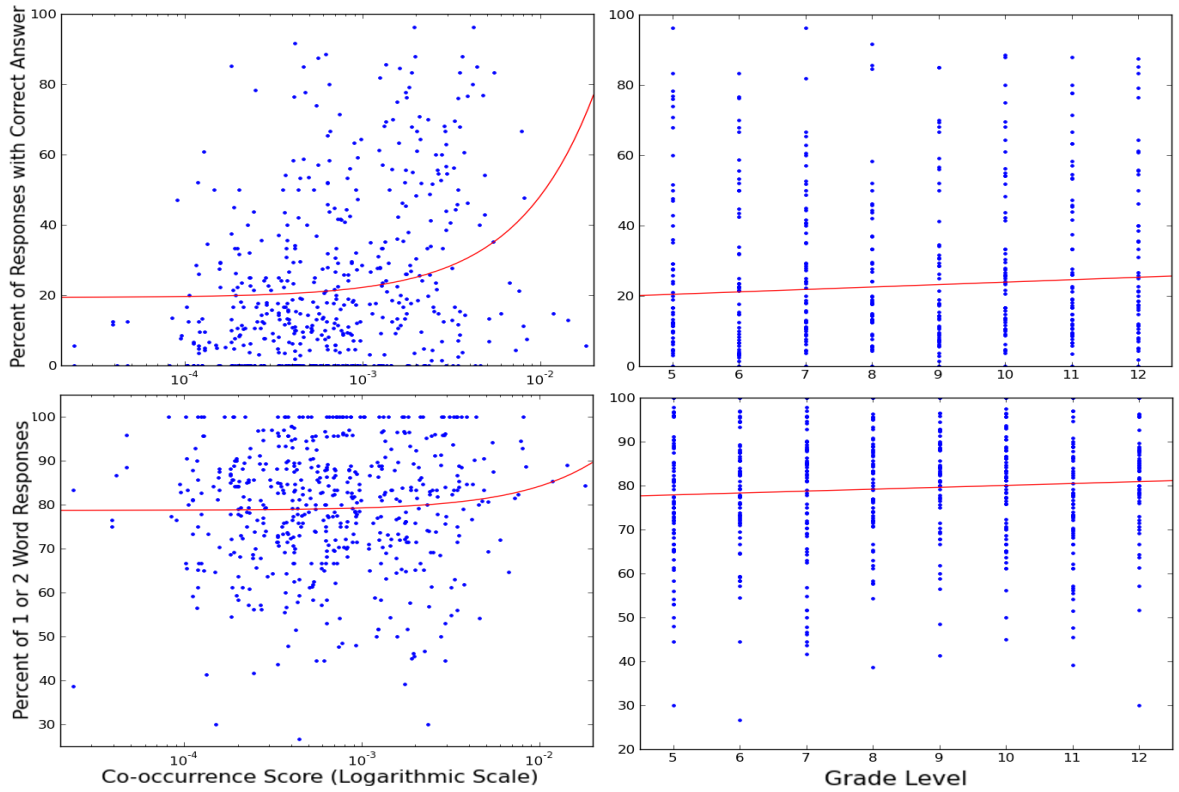

Figure (3): Scatter plots of all sentences with cloze quality measure as y-axis, and cloze quality estimator as x-axis. The linear regression of each distribution is shown.

one or two words. The linear regression shows there is a small (statistically insignificant) correlation.

The top-right plot shows Cloze Easiness (y-axis) per grade level (x-axis). The bottom left shows context restriction (y-axis) as a function of grade level. In both cases linear regressions here also show small, statistically insignificant positive correlations.

The lack of significant correlations for three out of four combinations of measures and estimators is not grounds to dismiss these measures. Across all sentences, the measure of context restriction is highly variant, at 47.9%. This is possibly the result of the methodology; in an attempt to avoid biasing the AMT workers, we did not specify the desirable number of answers. This led to many workers interpreting the task differently.

In terms of maximum grade level, the lack of a significant correlation with context restriction does not absolutely refute Finn (1978)'s hypothesis. Finn specifies that semantic transfer features should be in "lexical scope" of a blank. A clear definition of "lexical scope" was not presented. We generalized scope to mean proximity within a fixed contextual window size. It is possible that a more precise definition of "lexical scope" will provide a stronger correlation of reading level and context restriction.

5.1 Expert Validation

Finally, while we have shown a statistically significant positive correlation between co-occurrence scores and Cloze Easiness, we still need to demonstrate that Cloze Easiness is a valid measure of cloze quality. To do so, we selected the set of 20 sentences that ranked highest by co-occurrence score and by Cloze Easiness to submit to expert evaluation. Due to overlap between these two sets, choosing distinct sentences for both would require choosing some sentences ranked below the top 20 for each category. Accordingly, we chose to submit just one set based on both criteria in combination.

Along with these 20 sentences, as controls, we also selected two more distinct sets of 20 sentences: one set of sentences measuring most highly in context restriction, and one set most highly estimated by maximum grade level.

We asked a former English teacher to read each open cloze, without the key, and rate, on a five point Likert scale, her agreement with the statement "*This is a very good fill-in-the-blank sentence.*" where 1 means strong agreement, and 5 means strong disagreement.

		Expert evaluation on 5-point Scale	
		Mean	Standard Deviation
20 best sentences as determined by:	Cloze Easiness and co-occurrence score	2.25	1.37
	Context restriction	3.05	1.36
	Maximum grade level	3.15	1.2

Table (2): Mean ratings for each sentence category.

The results in Table (2) show that, on average, the correlated results of selecting sentences based on Cloze Easiness and co-occurrence score are in fact rated more highly by our expert as compared to sentences selected based on context restriction, which is, in turn, rated more highly than sentences selected by maximum grade level. Using a one-sample t-test and a population mean of 2.5, we find a p-value of .0815 for our expert's ratings.

6 Conclusion

We present a multi-step filter-based paradigm under which diverse estimators of cloze quality can be applied towards the goal of full automation of cloze task generation. In our implementation of this approach sentences were found for a set of keys, and then filtered by maximum length and likelihood of well-formedness. We then tested combinations of two estimators and two experimental measures of cloze quality for the next filtering step.

We presented an information-theoretical basis for the use of reading level as a novel estimator for cloze quality. The hypothesis that maximum grade level should be correlated with context restriction was not, however, shown with statistical significance. A stronger correlation might be shown with a different experimental methodology and a more refined definition of lexical scope.

As an alternative to expert evaluation of cloze quality, we investigated the use of non-expert workers on AMT. A statistically significant correlation was found between the co-occurrence score of a sentence and its experimental measure of Cloze Easiness. This is evidence that crowdsourcing techniques agree with expert evaluation of co-occurrence scores in past studies.

To gain further evidence of the validity of these experimental results, sentences selected by a composite filter of co-occurrence score and Cloze Easiness were compared to sentences selected by context restriction and reading level. An expert evaluation showed a preference for sentences selected by the composite filter.

We believe that this method of cloze task selection is promising. It will now be tested in a real learning situation. This work contributes insight into methods for improving technologies such as intelligent tutoring systems and language games.

References

Alderson, J. C. (1979). The Cloze Procedure and Proficiency in English as a Foreign Language. *TESOL Quarterly*, *13*(2), 219-227. doi: 10.2307/3586211.

Bachman, L. F. (1982). The Trait Structure of Cloze Test Scores. *TESOL Quarterly*, *16*(1), 61.

Brown, J., & Eskenazi, M. (2004). Retrieval of Authentic Documents for Reader-Specific Lexical Practice. In *InSTIL/ICALL Symposium* (Vol. 2). Venice, Italy.

Collins-Thompson, K., & Callan, J. (2005). Predicting reading difficulty with statistical language models. *Journal of the American Society for Information Science and Technology*, *56*(13), 1448-1462.

Dale, E. (1965). Vocabulary measurement: Techniques and major findings. *Elementary English*, *42*, 395-401.

Dale, E., & Chall, J. S. (1948). A Formula for Predicting Readability: Instructions. *Educational Research Bulletin*, *Vol. 27*(2), 37-54.

Finn, P. J. (1978). Word frequency, information theory, and cloze performance: A transfer feature theory of processing in reading. *Reading Research Quarterly*, *13*(4), 508-537.

Francis, W. N. & Kucera, H. (1979). Brown Corpus Manual, *Brown University Department of Linguistics.* Providence, RI

Heilman, M., Collins-Thompson, K., & Eskenazi, M. (2008). An Analysis of Statistical Models and Features for Reading Difficulty Prediction. *3rd Workshop on Innovative Use of NLP for Building Educational Applications*. Assoc. for Computational Linguistics.

Higgins, D. (2006). Item Distiller: Text retrieval for computer-assisted test item creation. ETS, Princeton, NJ.

Hoshino, A., & Nakagawa, H. (2005). A real-time multiple-choice question generation for language testing – a preliminary study–. In *2nd Workshop on Building Educational Applications Using NLP* (pp. 17-20). Ann Arbor, MI: Association for Computational Linguistics.

Jongsma, E. (1980). Cloze instructional research: A second look. Newark, DE: International Reading Association. Urbana, IL.

Kincaid, J., Fishburne, R., Rodgers, R., & Chissom, B. (1975). Derivation of new readability formulas for navy enlisted personnel. *Research Branch Report*. Millington, TN.

Klein, D. & Manning, C. (2003). Accurate Unlexicalized Parsing. (pp. 423-430) In *Proceedings of the 41st Meeting of the Assoc. for Computational Linguistics.*

Lee, J., & Seneff, S. (2007). Automatic Generation of Cloze Items for Prepositions. *Proceedings of In*.

Liu, C., Wang, C., Gao, Z., & Huang, S. (2005). Applications of Lexical Information for Algorithmically Composing Multiple-Choice Cloze Items. In *Proceedings of the 2nd Workshop on Building Educational Applications Using NLP* (p. 1–8). Ann Arbor, MI: Association for Computational Linguistics.

Open American National Corpus (2009) *americannationalcorpus.org/OANC/*

Perfetti, C., & Hart, L. (2001). *Lexical bases of comprehension skill*. (D. Gorfein) (pp. 67-86). Washington D.C.: American Psychological Association.

Petersen, S. E., & Ostendorf, M. (2006). Assessing the reading level of web pages. In *ICSLP* (Vol. pages, pp. 833-836).

Pino, J., Heilman, M., & Eskenazi, M. (2008). A Selection Strategy to Improve Cloze Question Quality. In *Proceedings of the Workshop on Intelligent Tutoring Systems for Ill-Defined Domains*.

Pito, R. (1994). tgrep README *www.ldc.upenn.edu/ldc/online/treebank/README.long*

Resnik, P., & Elkiss, A. (2005). The Linguist's Search Engine: An Overview. *Association for Computational Linguistics*, (June), 33-36.

Shannon, C. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, *27*, 379–423, 623–656.

Taylor, W. L. (1953). Cloze procedure: a new tool for measuring readability. *Journalism Quarterly*, *30*, 415-453.