Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 7094–7101
Marseille, 11–16 May 2020

Improving the Production Efficiency and Well-formedness of Automatically-Generated Multiple-Choice Cloze Vocabulary Questions

Ralph L. Rose
Waseda University Faculty of Science and Engineering
Tokyo, Japan
rose@waseda.jp

Abstract

Multiple-choice cloze (fill-in-the-blank) questions are widely used in knowledge testing and are commonly used for testing vocabulary knowledge. Word Quiz Constructor (WQC) is a Java application that is designed to produce such test items automatically from the Academic Word List (Coxhead, 2000) and using various online and offline resources. The present work evaluates recently added features of WQC to see whether they improve the production quality and well-formedness of vocabulary quiz items over previously implemented features in WQC. Results of a production test and a well-formedness survey using Amazon Mechanical Turk show that newly-introduced features (Linsear Write readability formula and Google Books NGrams frequency list) significantly improve the production quality of items over previous features (Automated Readability Index and frequency list derived from the British Academic Written English corpus). Items are produced faster and stem sentences are shorter in length without clear degradation in their well-formedness. Nearly 90% of such items are judged well-formed, surpassing the rate of manually-produced items.

Keywords: multiple-choice cloze, vocabulary testing, automated test creation, readability formula, frequency list

1. Introduction

A standard part of any language teaching curriculum is a focus on vocabulary study. Any program of vocabulary study consists of several components: some specification of the target lexical items, the order of their presentation to learners, presentation and study methods, and means of evaluation to gauge learners' mastery of the vocabulary. Although these components are interdependent, the present paper is focused on the last of these: means of evaluation. There are many ways of evaluating learners' mastery of vocabulary items. Lexical variation in a writing sample could be measured. Alternatively, fill-in-the-blank (hereafter, cloze) items may be used to evaluate learners' productive knowledge of best-fitting words. The most reliable approach with cloze items is likely free-response in which learners must think of the most suitable word(s) from their own lexical knowledge. But in large-scale testing situations, this may be impractical because checking answers may be time-consuming as the checker must make decisions about unexpected responses that might in fact be suitable completions. To avoid these problems a common approach is to check vocabulary knowledge with multiple-choice cloze items. These can be checked very quickly and reliably by hand, or nearly immediately if the evaluation is performed through an online interface such as a web browser. But the burden with multiple-choice cloze questions is the time it takes to prepare them in advance. Fortunately, a number of automated applications exist that may produce such vocabulary questions. The present paper is a report on recent developments in one of these, Word Quiz Constructor (WQC: Rose, 2014a, 2014b, 2016). Earlier work demonstrated that WQC can produce multiple-choice cloze vocabulary quiz items that are comparable in well-formedness to manually-produced items and at a modestly faster rate.

The present report outlines improvements made to WQC to yield items that are produced faster and are more often well-formed and the optimal configuration of resources and settings to produce items. The remainder of the paper is organized as follows. The following section reviews earlier work on multiple-choice questions for vocabulary knowledge evaluation and then describes WQC in detail including its new features. After that, an experiment to evaluate the production speed and effectiveness of the new features of WQC is described followed by the results of that experiment. Finally, the results are discussed in terms of what they suggest about the optimal use of WQC for production of vocabulary quiz items and future development plans for the application are described in detail.

2. Background

2.1 Automatic production of language testing questions

A typical multiple-choice cloze item consists of the following components, illustrated in (1) below. A *stem* sentence containing one or more blanks which represent a sequence of words removed from the sentence. Although multi-word sequences and multiple blanks are certainly possible, the simplified case dealt with in this paper is single blanks filled by a single word. Therefore, hereafter, discussion of multiple-choice cloze items will assume this simple configuration. The word which fits in the blank is called the *key*. After the stem, several answer *options* are shown, including the key and some *distractor* options which do not optimally fit in the blank relative to the key. Hence, in (1), the key is "corpus", and the distractors are "batch", "package", and "wardrobe".

(1) A _____ is a collection of writings.
a. batch b. package c. corpus d. wardrobe

Many computer application systems exist which are designed to produce multiple-choice cloze question items for testing linguistic skills (e.g. Goto et al, 2010; Kunechika et al, 2003; Mitkov et al, 2006, 2009; Pino et al, 2008; Sumita et al, 2005). Many of these systems depend on online and offline resources such as the British National Corpus, Wikipedia, Google services, and Wordnet. A generalized procedure for the construction of such items is as follows, beginning with a source text.

1. Using various statistical methods, and relying on online or offline resources, select a sentence from the

text that represents an important concept or uses a crucial word. Make this sentence the stem.
2. Select one word from the sentence as the key.
3. Select a number of words which are semantically similar to the key but which are (relatively) implausible completions (due to, say, grammatical fit). Use online or offline resources to evaluate semantic similarity and plausibility.
4. Finalize the item and output in a desired format.

This basic procedure has been used to successfully create quiz questions that test text comprehension (e.g. citations in above paragraph), or to evaluate vocabulary knowledge (e.g. Aist, 2001; Brown et al, 2005; Coniam, 1997; Heilman and Eskenazi, 2007).

These applications, while broadly useful, are not suitable to all vocabulary testing needs. For instance, one commonly used approach to vocabulary teaching is the regular provision of vocabulary lists for the learner to study, followed by quizzes focusing on just the items on those lists (cf. Brown and Perry, 1991; Khoii and Sharififar, 2013; Sagarra and Alba, 2006). In this context, there is no source text to provide as an input to the question generation procedure. Instead, the procedure must start with the selection of a target word from the vocabulary list. Then, a stem sentence needs to be chosen from some suitable outside source. None of the above systems are capable of this (although systems by Lee et al, 2013 and Liu et al, 2005 come close). Furthermore, this context places certain constraints on question generation. For instance, when selecting distractor options, the options must come from the same list as the key. If they are from a different list or drawn from external sources, then the vocabulary question is no longer a test of learners' knowledge of the words in the list but rather their simple recognition of which words are or are not in the list being tested.

2.2 Word Quiz Constructor

2.2.1 Basic architecture

WQC was designed and built to meet the need for automatic word quiz generation focused on specified word lists. It was also built with the idea of mass generation of quizzes to enable its use in a large-scale language program (cf. Abu-Alhija, 2007; Fulcher and Davidson, 2007; Weir, 2005). Figure 1 illustrates in flow-chart format the production of a single multiple-choice cloze item.

WQC starts by randomly selecting a word (e.g. 'positive) from one or more specified sublists of the Academic Word List (AWL: Coxhead, 2000). It then retrieves a random sentence from Wikipedia containing the word ('It was released to positive reviews on April 11, 2003.') and checks the language-wide frequency of the trigram containing the target word at the center ('to positive reviews'). If this frequency does not exceed a specified threshold, a new random sentence will be retrieved and the trigram frequency will be evaluated. This process continues until a suitable high-frequency context is found (or until a specified limit is reached and WQC gives up on the currently selected word and begins over again with a new randomly-selected word). Before continuing, WQC also checks that the stem sentence has a readability level that does not exceed a user-specified level of difficulty.

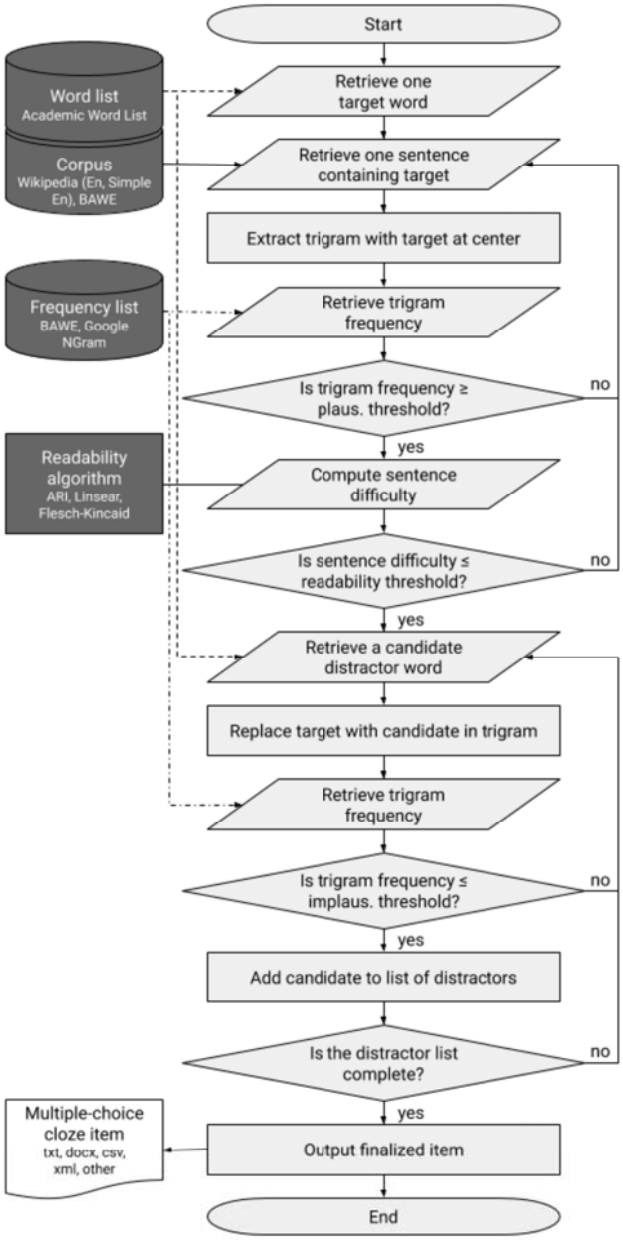

Figure 1: Flow chart for the production of a single multiple-choice cloze item by Word Quiz Constructor.

Once a suitable stem sentence and a suitable high-frequency context for the key word is found, WQC searches the same specified AWL sublist(s) for distractor options that are not suitable for the context. It does this by substituting the key word in the original trigram with the candidate distractor (e.g. 'to acquiring reviews') and checks if its frequency is below a specified threshold. When a specified number of distractors are found (e.g. three, as in sample (1) above), the item is finalized and output in a desired format. WQC is capable of outputting in plain text and rich text format (i.e. Microsoft docx) as well as in data formats (e.g. csv, xml) for import into online testing environments such as Moodle.

In addition to multiple-choice cloze questions, WQC can also produce synonym questions (e.g. "Which word is closest in meaning to the highlighted word?") and free-response cloze questions (similar to multiple-choice cloze questions but with no options, allowing the learner to

freely write any suitable option they know). These other questions are not evaluated in the present report and are not discussed further. Their details will be reported in later papers.

WQC is programmed in Java and runs in console mode. One "run" of the program will create one *quiz*, which has a user-specified structure: for example, 10 multiple-choice cloze items, 5 synonym items, and 5 free-response cloze items.

2.2.2 Resources

WQC depends on a number of online and offline resources for its functionality. This section describes these resources and the user-defined parameters that go with each. It also notes which resources have been newly integrated and which are being tested in the present work.

Word list WQC retrieves words from AWL (Coxhead, 2000). This list consists of high-frequency words in academic documents and lists 570 word families. A word family consists of a head word and its derived variants (e.g. the "analyze" family consists of headword "analyze" and other members "analyzed", "analyzing", "analytic", "analyzable", etc.). These families are organized into 10 sublists known as Sublists 1 to 10 and are in decreasing order of frequency (thus, Sublist 10 words are of lowest frequency). A WQC user may specify one or more sublists to retrieve words from. WQC will retrieve a random word and attempt to create an item by retrieving a stem and selecting options.

Corpus WQC is capable of retrieving stem sentences from three different corpora: the British Academic Written English (BAWE) Corpus (Gardner and Nesi, 2012), the default English version of Wikipedia (EW: en.wikipedia.org) and the Simple English version of Wikipedia (SEW: simple.wikipedia.org). While the texts in all of these corpora are written in an academic style and tone, the articles in SEW are written to be simpler, using "only the 1,000 most common and basic words in English" and "simple grammar and shorter sentences" (Wikipedia contributors, 2019). When constructing vocabulary quiz items for English as a second (ESL) or foreign language (EFL) learners, SEW might be preferred. In fact, earlier studies with WQC showed that EW was more reliable for constructing items than BAWE (Rose, 2014a) and that SEW was more reliable for constructing items than EW (Rose, 2016) when the aim is to create items with easier-to-read stems (i.e. lower grade level as computed by readability algorithms; discussed in detail below).

Frequency list WQC relies on a trigram frequency list for two purposes. First, after retrieving a candidate stem and extracting the trigram with the key word at the center, the frequency of the trigram is retrieved from the frequency list. If the frequency meets or surpasses a user-specified plausibility threshold, then the context is regarded as a high plausibility context for the key word and the candidate sentence is accepted as the stem for the current item. Thereafter, each candidate distractor word is substituted into the trigram for the key word and the frequency of the modified trigram is retrieved from the frequency list. If the frequency is no higher than a user-specified implausibility threshold, then the context is regarded as a low plausibility context for the option word and is accepted as a distractor option. In this sense, the plausibility threshold acts like a high-pass filter and the implausibility threshold acts like a low-pass filter.

Because the earliest version of WQC was aimed at producing items to test AWL words and the produced quizzes were intended for use in an academic context (a course in academic reading in English at a university-level science and engineering program), the initially selected frequency list was the set of trigrams drawn from BAWE. Although BAWE is not a small corpus at 6.5 million words, neither is it a very large corpus. Therefore, the trigram list is not as extensive as those derived from larger corpora. As a result, it was found that the plausibility threshold could not be set much higher than 2 or very few contexts would ever pass the plausibility standard. Furthermore, although the implausibility threshold was set at 0, the somewhat lower coverage of the trigram frequency list meant that there could more easily be trigrams with 0 frequency that are actually plausible but just not attested in the BAWE corpus (i.e. false negatives). This leads to the possibility of distractor options that are, in fact, suitable completions for the cloze item in addition to the key word.

In order to address this perceived shortcoming, the present version of WQC now optionally uses the English (US) trigram list from Google Books NGrams (GBNG, Version 2: Michel et al, 2011). The trigram list is accessed online via the PhraseFinder API (Trenkman, 2019) and includes 419 million trigrams. With some manual experimentation it was found that a plausibility threshold of 100 and an implausibility threshold of 0 was sufficient to allow WQC to process items smoothly.

In the present work, these two frequency lists—BAWE and GBNG—are compared for their effectiveness in constructing multiple-choice cloze vocabulary questions.

Readability algorithm WQC uses a readability algorithm to estimate the difficulty of the candidate stems. Ideally, the stem sentence should be easy to read and understand so that the learner is merely required to discover which word optimally fills in the gap based on their knowledge of the word. If the stem sentence is very difficult, then many language learners may struggle to choose the right option, but not necessarily because they do not know the meaning of the key word. But for highly advanced learners, difficult stem sentences may be useful to test the depth of their vocabulary knowledge. In any case, controlling readability is useful for adapting WQC output to learners' levels.

WQC has made use of the automated readability index (ARI: Smith and Senter, 1967) as shown in Figure 2(a). This is a very easy to compute index which needs nothing more than the text itself to compute its value. The result is an estimate of the US grade level reading skill required to understand the sentence: That is, 0 to 6 being elementary school, 7 to 12 being junior high and senior high school, and greater than 12 being college and beyond.

(a) $$ARI = 4.71\left(\frac{characters}{words}\right) + 0.5\left(\frac{words}{sentences}\right) - 21.43$$

$$LW_{provisional} = \frac{(num\ of\ 1,2\ syll\ words) + 3(num\ of\ 3+syll\ words)}{num\ sentences}$$

(b) $$If\ LW_{provisional} > 20, LW = \frac{1}{2} LW_{provisional}$$

$$Else\ if\ LW_{provisional} \leq 20, LW = \frac{1}{2} LW_{provisional} - 1$$

(c) $$FK = 0.39\left(\frac{words}{sentences}\right) + 11.8\left(\frac{syllables}{words}\right) - 15.59$$

Figure 2: Readability formulas used by WQC: (a) Automated Readability Index (ARI), (b) Linsear Write formula (LW), and (c) Flesch-Kincaid grade level formula (FK).

While ARI is simple and fast, there has been some concern that it is not capturing reading difficulty reliably for ESL/EFL learners. Therefore, the present version of WQC integrates two new more sophisticated measures of readability for comparison purposes. They are the Linsear Write formula (LW: Klare, 1974) and the Flesch-Kincaid grade level formula (FK: Kincaid et al, 1975), shown in Figure 2(b) and Figure 2(c), respectively. Crucially, these are both dependent on lexico-phonological information via syllable count. The syllable information about words is obtained via reference to the CMU Pronouncing Dictionary (CMUdict: Carnegie Mellon University, 2019). In the present work, these three readability formulas—ARI, LW, and FK—are compared for their effectiveness in constructing multiple-choice cloze vocabulary questions.

3. Experiment

In order to examine the effectiveness of WQC with its new features, a two-part experiment was performed: timed generation of multiple quizzes in various configurations and then a crowd-sourced evaluation of the well-formedness of a sample of these items by native English speakers. This section describes this experiment and the results.

3.1 Method

The experiment is focused on comparing the production of multiple-choice cloze items in various configurations of source corpus, frequency list, and readability formula. These three factors are manipulated within the following levels.

- Source corpus: English Wikipedia (EW), Simple English Wikipedia (SEW)
- Frequency list: British Academic Written English corpus (BAWE), Google Books NGrams (GBNG)
- Readability formula: Automated Readability Index (ARI), Linsear Write (LW), Flesch-Kincaid grade level (FK)

Although WQC is capable of using BAWE as a source corpus, as noted above, previous work has shown that BAWE is clearly less effective for producing items than EW. Therefore, in the present work, only EW and SEW are used for comparative purposes.

First, WQC was made to produce 10 quizzes consisting of 10 multiple-choice cloze items with four answer options each (i.e. as in example (1) above) based on random words taken from AWL Sublists 1 and 2. This was done in all twelve configurations of the three main factors (corpus x frequency list x readability formula). Hence, 100 items total were produced in each configuration for a total of 1,200 items. Because the size of the trigram frequency lists differ so much, it was impossible to use the same plausibility threshold for both (though the implausibility threshold for both was the same: 0). Based on previous experience with WQC the plausibility thresholds were set at 2 for BAWE and 100 for GBNG. Finally, the target learner was assumed to be university level ESL/EFL learners—a typical level at which AWL words are a pedagogical focus. Therefore, in order to assure that the reading level was not too difficult, a maximum readability threshold of 12 was set. Note that all three readability formulas reference the same scale (US grade level), so the same threshold value was set no matter which formula was used.

In order to evaluate the well-formedness of the items produced, a sample of eight items in each configuration were selected pseudo-randomly to create 12 groups with comparable mean readability grade level. Furthrmore, 16 check items were included with the actual test items: 8 well-formed and 8 not well-formed items. The well-formedness of the check items were independently verified in earlier work (Rose, 2014a). In addition to these 96 items plus 16 check items, 16 items produced manually by an experienced EFL teacher were included for comparison.

These 128 items were presented to native speakers of English (self-reported nativeness) through the Amazon Mechanical Turk crowd-sourcing system. Items were presented with their keys already selected and workers were asked to judge whether the item was a well-formed vocabulary question. At the start of the task, instructions explained that well-formedness included having a stem sentence that was not more difficult than the word being tested and having only a single best option as highlighted. Several possible reasons for non-wellformedness were given and two sample items (one well-formed, the other not well-formed) were also shown and explained in depth.

The complete task for each worker therfore included reading a consent form, reading the instructions, judging the well-formedness of 96 WQC + 16 manual + 16 check = 128 items (order of all items randomized for each worker), and responding to two demographic questions (native English speaker status yes/no, highest completed education level). Each worker was paid US$7.50 for the successful completion of the task.

Workers were recruited within the Mechanical Turk system and limited only to those workers who met the following three qualifications: (a) have already completed 10,000 tasks (i.e. Human Intelligence Tasks, or HITs), (b) have a 99 % or greater acceptance rate on previously completed tasks, and (c) are located in either Canada or USA. The first two qualifications help to increase the

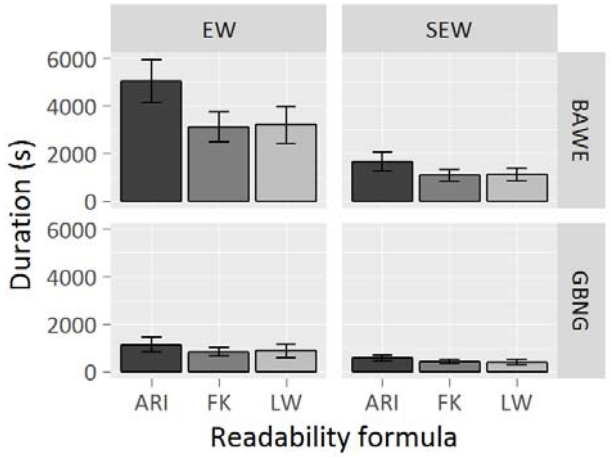

Figure 3: Timing information of the production of multiple-choice cloze items by WQC. Error bars indicate 95 % confidence intervals (in all graphs).

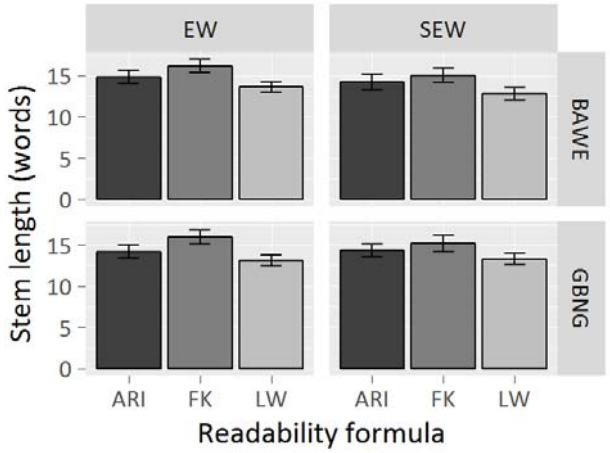

Figure 4: Stem length in words of the multiple-choice cloze items produced by WQC.

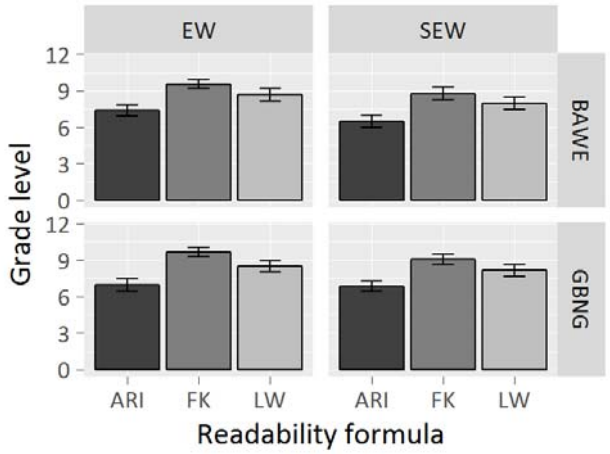

Figure 5: Grade level of stems of the multiple-choice cloze items produced by WQC.

likelihood that those who do the task are careful workers who will read the instructions and do the task conscientiously. The last qualification is intended to minimize variation from English varietal differences.

3.2 Results

The results from the two parts of the experiment (production, well-formedness judgments) are presented below. The statistical analyses were performed in `R` (version 3.3.2) using linear regression modeling with `lm` and mixed effects modeling with `lme` (package `nlme`, version 3.1-128) and using $\alpha = 0.05$. Results are summarized in Figures 3 to 7 in which error bars indicate 95% confidence intervals.

3.2.1 Production: timing

The production timing results are summarized in Figure 3, showing the duration of time to produce one 10-item quiz. These results show that SEW items are produced faster (in about one-third the time) than EW items [$t(108) = 12.3$, $p < .001$]. Furthermore, using FK and LW to estimate readability leads to items that are produced somewhat faster (in about two-thirds the time) than those using ARI [$t(108) = 6.7$, $p < .001$]. Finally, using GNBG as a frequency list is faster than using BAWE [$t(108) = 14.2$, $p < .001$]. Overall, the most optimal configuration for producing items quickly is by drawing items from SEW and estimating their readability using FK or LW and determining plausible (i.e. high-frequency) contexts using GBNG as a frequency list. In this configuration, a single item is produced in 43 seconds (FK) and 41 seconds (LW) on average.

3.2.2 Production: stem length

The stem length results are summarized in Figure 4, showing the mean length of selected stems in each configuration. These results show only one clear effect: Items produced using LW as a readability measure yield slightly shorter stem sentences than the other measures [$t(1188) = 2.0$, $p < .05$]. On average, the LW items are about two words shorter than the ARI and FK items.

3.2.3 Production: grade level

The grade level results are summarized in Figure 5, showing the estimated reading grade level of the stem sentences (according to the respective readability formula used to produce the item). Results show that items with stems from SEW are estimated to be a little more readable than those with stems from EW [$t(1188) = 2.7$, $p < .01$] by about one-half of a grade level on average. However, the largest effect seems to be with items produced with the ARI readability formula: These are significantly more likely to be easier to read than those produced with the FK or LW formulas [$t(1188) = 3.7$, $p < .001$]—almost two grade levels lower on average.

3.2.4 Well-formedness judgments

The task was successfully completed by 41 workers. However, a minimum of 80% accuracy on the check items was set as a limit for inclusion in the subsequent analysis. As a result the following analyses are based on $n = 37$ workers. In the Mechanical Turk system, there is no way to know precisely how long workers took to complete the task but the fastest workers completed the task in just under 30 minutes (this compared to an advertised estimated completion time of "less than one hour"). The interrater agreement of these 37 workers on the 16 check items was very high: Fleiss' $\kappa = 0.872$ (note : for all 41 workers, $\kappa = 0.766$).

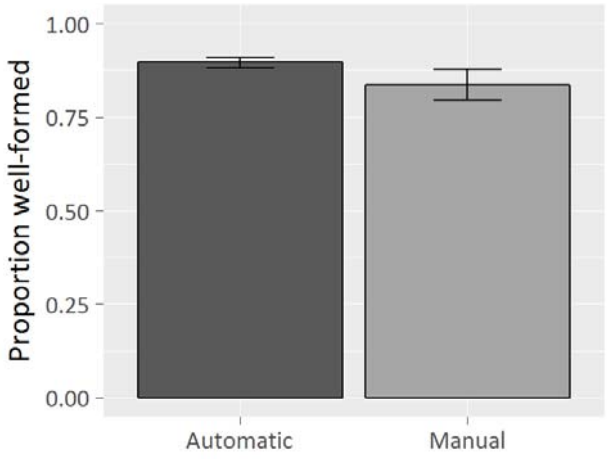

Figure 6. Proportion of items produced by WQC vs. manually-produced items judged well-formed.

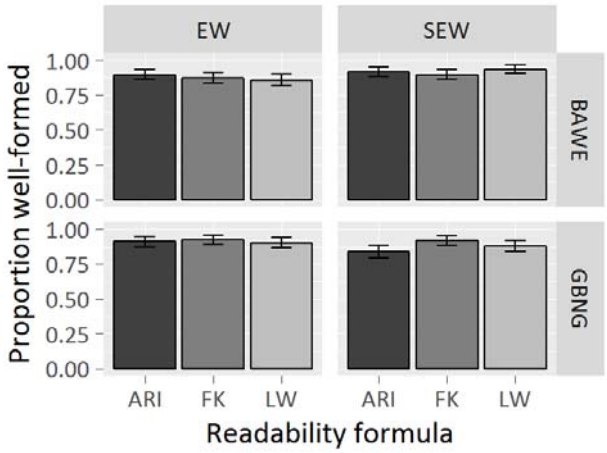

Figure 7. Proportion of items produced by WQC which were judged well-formed.

The well-formedness judgments for manually-produced versus automatically-produced items are summarized in Figure 6, showing the proportion of items in each category judged well-formed by the Mechanical Turk workers. Mixed effects modeling (with participants and items as random effects) shows that the automatically-produced items are more often judged well-formed than the manually-produced items [$t(2885) = 4.3$, $p < .001$] with about 90% of automatic items judged well-formed compared to about 84% for manual items.

The well-formedness judgments for automatic items is broken down and summarized in Figure 7, showing the proportion judged well-formed in each configuration. Interestingly, there is only one significant difference: There is an interaction between corpus and frequency list [$t(3314) = 2.6$, $p < .01$] showing that the well-formedness judgments of items produced from SEW and using the GBNG frequency list are slightly lower than others. However, this difference is small (e.g. 87% judged well-formed for SEW/GBNG items versus 91% for either SEW/BAWE items or EW/GBNG items. Furthermore, the marginal R^2=0.01 while the conditional R^2=0.88 suggesting that most of the difference is more likely explained by random variation.

4. Discussion

This paper has examined the most recently introduced features of Word Quiz Constructor (WQC): the use of Google Books NGrams (GBNG) as a richer frequency list and the use of more sophisticated readability formulas—Flesch-Kincaid grade level (FK) and Linsear Write (LW) and evaluated them against already existing features to see what, if any improvements they offer in production and well-formedness of items produced by WQC. Results show that all configurations of corpus, frequency list, and readability formula produce English multiple-choice cloze vocabulary quiz items that are judged comparably well-formed by native English readers. Furthermore, these items are superior (by a small but significant margin) to manually-produced items.

The main difference between the various configurations arises at the production stage. The optimal configuration overall would seem to be to use Simple English Wikipedia (SEW) as a source corpus for the stem sentence and GBNG as the frequency list because these produce items the fastest, while using LW as the readability formula because it tends to produce shorter stems. This configuration does have slightly lower well-formedness judgments, but not appreciably lower.

If a slightly higher well-formedness rate is desired, an alternative may be to retrieve stems from EW rather than from SEW. However, this is at the cost of roughly doubling production time. If creating only one short quiz, this may be negligible. But if creating a large number of quizzes en masse for a large-scale program, the time increase may be unacceptable.

If minimizing production time is not a priority at all, then the Automated Readability Index (ARI) may be used as readability formula instead of LW. At the cost of a significantly longer production time and slightly longer stem sentences. It will produce stems that have a lower reading grade level. However, it is possible that some of this grade level advantage is merely an artifact of the simplicity of the ARI formula. Note that three-character acronyms (IMF, DNA, CMU) and chemical formulas (CO_2, H_2O) would be seen by ARI the same as any three-character words (and, but, see) but would be counted as three-syllable words (so-called "hard" words in the LW scheme) by both the LW and FK readability formulas. Thus, it is possible that the ARI grade level advantage is partially illusory. The fact that the ARI items are not judged well-formed more often than those of LW and FK would seem to support this conjecture.

One final issue worth discussing is the fact that the manual items did not themselves receive the highest well-formedness rating. This reflects the fact that even items produced manually by an experienced language teacher will not be perfect—perhaps creating only 90 well-formed items out of 100 tries. What the present work shows is that WQC can achieve that rate plus slightly more, freeing the teacher to focus on other tasks while WQC does its job. Afterward, the teacher can then just focus on identifying and fixing the remaining 10.

5. Future work

The present work has evaluated well-formedness in English multiple-choice cloze vocabulary questions using native English speakers in the Amazon Mechanical Turk system. However, it would be useful to confirm that the judgments comport with the intuitions of experienced teachers by replicating the study with such teachers, particularly those with strong vocabulary test creation experience. This is the next anticipated step in the continuing development and evaluation of WQC.

Besides this, the three readability formulas compared in the present work are, in fact, not fully suited to the task they are put to in WQC. They are designed to work with longer texts rather than with single sentences, and research shows that their reliability increases as the evaluated text length increases (cf. Zhou, Jeong, and Green, 2017). Thus, a search for a more reliable formula for judging the readability of single sentences continues.

Finally, while WQC is a stable application, it has not been released publicly. Therefore, preparations are underway for a public release so that other language testers may use it to generate uniform vocabulary quizzes for their learners. In the meantime, researchers or teachers who wish to obtain a pre-release copy of WQC may contact the author directly. Developers who are interested in seeing the underlying code may also request the code archive from the author.

Acknowledgements

This research has been partially funded by a Waseda University Grant for Special Research Projects (Project #2019C-236).

6. Bibliographical References

Abu-Alhija, F.N. (2007). Large-scale testing: Benefits and pitfalls. *Studies in Educational Evaluation*, 33(1):50–68.

Aist, G. (2001). Towards automatic glossarization: automatically constructing and administering vocabulary assistance factoids and multiple-choice assessment. *International Journal of Artificial Intelligence in Education*, 12:212–231.

Amazon Mechanical Turk web site. https://www.mturk.com.

Brown, J., Frishkoff, G. and Eshkenazi, M. (2005). Automatic question generation for vocabulary assessment. In Raymond J. Mooney (Conference Chair), *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 819–826, Vancouver, British Columbia, Canada, October. Association for Computational Linguistics (ACL).

Brown, T.S. and Perry, F.L. (1991). A Comparison of Three Learning Strategies for ESL Vocabulary Acquisition. *TESOL Quarterly*, 25(4):655–670.

Coniam, D. (1997). A Preliminary Inquiry Into Using Corpus Word Frequency Data in the Automatic Generation of English Language Cloze Tests. *CALICO Journal*, 14(2-3):15–33.

Coxhead, A. (2000). A New Academic Word List. *TESOL Quarterly*, 34(2):213–238.

Fulcher, G. and Davidson, F. (2007). *Language testing and assessment*. London & New York: Routledge.

Gardner, S. and Nesi, H. (2012). A classification of genre families in university student writing. *Applied Linguistics*, 34(1):1–29.

Goto, T., Kojiri, T., Watanabe, T., Iwata, T. and Yamada, T. (2010). Automatic Generation System of Multiple-Choice Cloze Questions and its Evaluation. *Knowledge Management & E-Learning*, 2(3):210–224.

Heilman, M. and Eskenazi, M. (2007). Application of Automatic Thesaurus Extraction for Computer Generation of Vocabulary Questions. In Proceedings of Speech and Language Technology in Education (SLaTE), pages 65–68, Farmington, Pennsylvania, USA, October, International Speech Communication Association (ISCA) special interest group for Speech and Language Technology in Education.

Khoii, R. and Sharififar, S. (2013). Memorization versus semantic mapping in L2 vocabulary acquisition. *ELT Journal*, 67(2):199–209.

Kincaid, J.P., Fishburne, R.P., Rogers, R.L., & Chissom, B.S. (1975). Derivation of new readability formulas (automated readability index, fog count, and flesch reading ease formula) for Navy enlisted personnel. *Research Branch Report 8–75*. Chief of Naval Technical Training: Naval Air Station Memphis.

Klare, G.R. (1974). Assessing Readability. *Reading Research Quarterly*, 10(1):62–102.

Kunichika, H., Katayama, T., Hirashima, T. and Takeuchi, A. (2001). Automated question generation methods for intelligent English learning systems and its evaluation. In *Proceedings of the International Conference on Computers in Education (ICCE)*, pages 1117–1124, Seoul, Korea.

Lee, K., Kweon, S., Kim, H. and Lee, G. (2013). Filtering-based Automatic Cloze Test Generation. In P. Badin et al, editors, Proceedings of Speech and Language Technology in Education (SLaTE), pages 72–76, Grenoble, France, August. International Speech Communication Association (ISCA) special interest group for Speech and Language Technology in Education.

Liu, C., Wang, C., Gao, Z., and Huang, S. (2005). Applications of Lexical Information for Algorithmically Composing Multiple-Choice Cloze Items. In Jill Burstein and Claudia Leacock, editors, *Proceedings of the 2nd Workshop on Building Educational Applications Using NLP*, page 1–8, Ann Arbor, Michigan, USA, June. Association for Computational Linguistics (ACL).

Michel, J-B., Shen, Y.K., Aiden, A.P., Veres, A., Gray, M.K., The Google Books Team, Pickett, J.P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M.A., and Aiden, E.L. (2011). Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*, 331(6014):176–182.

Mitkov, R., Ha, L.A., and Karamanis, N. (2006). A computer-aided environment for generating multiple-choice test items. *Natural Language Engineering*, 12 (2):177–194.

Mitkov, R., Ha, L.A., Varga, A., and Rello, L. (2009). Semantic similarity of distractors in multiple-choice tests: extrinsic evaluation. In Roberto Basili and Marco Pennacchiotti, editors, *Proceedings of the EACL 2009 Workshop on GEMS: GEometical Models of Natural*

Language Semantics, pages 49–56, Athens, Greece, March. Association for Computational Linguistics (ACL).

Pino, J., Heilman, M., Eskenazi, M. (2008). A Selection Strategy to Improve Cloze Question Quality. In Vincent Aleven, Kevin Ashley, Collin Lynch, and Niels Pinkwart, editors, *Proceedings of the Workshop on Intelligent Tutoring Systems for Ill-Defined Domains*, pages 22–32, Jhongli, Taiwan, June. Association for Computational Linguistics (ACL).

Rose, R. (2014a). Automated vocabulary quiz creation using online and offline corpora. Poster presentation at Teaching and Language Corpora (TaLC) conference, Lancaster, UK, July.

Rose, R. (2014b). WQC: A tool for quick automatic word quiz construction. Oral presentation at the 17th World Congress of the International Association for Applied Linguistics (AILA), Brisbane, Australia, August.

Rose, R. (2016). Automatic Word Quiz Construction Using Regular and Simple English Wikipedia. In L. Gómez Chova, A. López Martínez, I. Candel Torres, editors, *Proceedings of the International Technology, Education and Development Conference (INTED)*, pages 8032–8040, Valencia, Spain, March. International Academy of Technology, Education, and Development (IATED).

Sagarra, N. and Alba, M. 2006. The Key Is in the Keyword: L2 Vocabulary Learning Methods With Beginning Learners of Spanish. *The Modern Language Journal*, 90(2):228–243.

Smith, E.A. and Senter, R.J. 1967. Automated Readability Index. Air Force Systems Command, Wright-Patterson Air Force Base, Ohio, USA. *AMRL-TR-6620*.

Sumita, E., Sugaya, F., and Yamamoto, S. 2005. Measuring Non-native Speakers · Proficiency of English by Using a Test with Automatically-Generated Fill-in-the-Blank Questions. In Jill Burstein and Claudia Leacock, editors, *Proceedings of the 2nd Workshop on Building Educational Applications Using NLP*, pages 61–68, Ann Arbor, Michigan, USA, June. Association for Computational Linguistics (ACL).

Weir, C.J. (2005). *Language Testing and Validation: An Evidence-based Approach*. Houndgrave, Hampshire, UK: Palgrave-Macmillan.

Wikipedia contributors, "Wikipedia:Simple English Wikipedia," Wikipedia, The Free Encyclopedia, https://simple.wikipedia.org/w/index.php?title=Wikipedia:Simple_English_Wikipedia&oldid=6729342 (accessed December 1, 2019).

Zhou, S., Jeong, H., and Green, P.A. (2017). How Consistent Are the Best-Known Readability Equations in Estimating the Readability of Design Standards? *IEEE Transactions on Professional Communication*, 60(1):97–111.

7. Language Resource References

Carnegie Mellon University. (2019). CMUdict: Carnegie Mellon University Pronouncing Dictionary (version 0.7). http://www.speech.cs.cmu.edu/cgi-bin/cmudict (accessed October 1, 2019).

Trenkman, M. PhraseFinder—Search millions of books for language use. https://phrasefinder.io (accessed October 1, 2019).

Wikipedia Contributors. Wikipedia (English). https://en.wikipedia.org (accessed October 1, 2019)

Wikipedia Contributors. Wikipedia (Simple English). https://simple.wikipedia.org (accessed October 1, 2019).