OPEN ACCESS

Vocabulary Learning and Instruction
ISSN 2981-9954
Volume 14, Number 1 (2025)
https://doi.org/10.29140/vli.v14n1.2079

Castledown

Evaluation of Automated Vocabulary Quiz Generation with VocQGen

Qiao Wang
Center for English Language Education, Faculty of Science and Engineering (CELESE), Waseda University
judywang.jp@gmail.com

Ralph L. Rose
Center for English Language Education, Faculty of Science and Engineering (CELESE), Waseda University
rose@waseda.jp

Ayaka Sugawara
Center for English Language Education, Faculty of Science and Engineering (CELESE), Waseda University
ayakasug@waseda.jp

Naho Orita
Center for English Language Education, Faculty of Science and Engineering (CELESE), Waseda University
orita@waseda.jp

Abstract

VocQGen is an automated tool designed to generate multiple-choice cloze (MCC) questions for vocabulary assessment in second language learning contexts. It leverages several natural language processing (NLP) tools and OpenAI's GPT-4 model to produce MCC items quickly from user-specified word lists. To evaluate its effectiveness, we used the first sublist in the Academic Word List (AWL) to generate 60 questions with VocQGen. Then we compared the quality of 60 autogenerated questions with 40 manually created ones through expert reviews and through pilot testing with 68 students. Expert review results indicate that automatically generated questions exhibit higher grammatical accuracy and clearer contexts in question stems. However, the tool occasionally produces distractors that are acceptable as correct responses. Pilot testing results show that in general

the number of correct responses is higher in autogenerated questions, indicating the less challenging nature of these questions. The study concludes that manual check is still required for questions generated by VocQGen and future work should focus on improving distractor effectiveness.

Keywords: VocQGen, multiple-choice cloze questions, automated question generation, GPT-4.

Introduction

Vocabulary training and assessment is a fundamental part of nearly every second language development program (Alqahtani et al., 2015; Nation, 2022; Schmitt, 1997). Alongside this, multiple choice cloze (MCC) questions (Hale et al., 1989; Taylor, 1953; Zhang et al., 2023) are a mainstay of knowledge testing and have been useful in second language proficiency testing for decades. When testing vocabulary knowledge, a typical MCC question would look as shown in (1), where a carrier sentence (stem) is given which contains the target word being tested (key) but replaced with a blank. Following the stem are several options to fill in the blank including the key as well as several (often three) distractor words which are sub-optimal completions for the stem.

(1) *The scientist conducted a thorough _____ of the data.*
a. summary b. analysis c. hypothesis d. experiment

When constructed well, MCC questions are a reliable means of testing vocabulary knowledge (cf., O'Reilly & Streeter, 1977). The manual creation of MCC questions for small-scale testing (e.g., individual classes or courses) is a somewhat time-consuming task, though feasible. That is, it may take a teacher 15–20 minutes to create a ten-item quiz. However, various factors may complicate this process (cf., Abu-Alhija, 2007; Fulcher & Davidson, 2007; Weir, 2005) such as the need for test security if quizzes are created for multiple classes or for repeated use from term to term. Manual creation may also lead to unconscious yet systematic biases either in choice of stem sentences or distractors as well as variation in difficulty level of MCC items. These factors may grow to be much more significant in large-scale testing situations where testing reliability would be desired across multiple sections of a course taught by many different teachers over many time periods.

A solution to this challenge is the creation of tools that can automatically generate MCC items for vocabulary testing. The present work describes one such tool called VocQGen (Vocabulary Quiz Generator) which stems from a long-term project on Vocabulary Training and Testing (VocaTT; see Rose et al., 2022) in which the authors have been engaged since 2021. VocQGen leverages OpenAI's ChatGPT-4 large language model to generate highly reliable and relatively consistent MCC items in minimal time from user-specified word lists (cf., Abdullah et al., 2022).

The next section of the paper reviews relevant related work and explains the fundamental features needed in such a generation tool. After that, we share the procedures and results of a validation test of generated items relative to

manually-produced items. The paper wraps up with a discussion of the results and outlines some necessary points for further development of VocQGen.

Background

Many applications have been reported to produce MCC questions such as those described in Goto et al. (2010), Kunichika et al. (2001), Mitkov et al. (2006), Mitkov et al. (2009), Pino et al. (2008), and Sumita et al. (2005). These applications are designed to take a reading passage as input and generate an MCC question based on one of the vocabulary words within the text. As such, the assessment purpose is tied together with reading comprehension.

While useful, this approach may not be applicable to contexts where learners are provided with sequential lists of words to study and then are assessed on their knowledge of each list (cf., Brown & Perry, 1991; Khoii & Sharififar, 2013; Sagarra & Alba, 2006). To serve these contexts, other applications have been developed such as Brown et al. (2005), Coniam (1997), Lee et al. (2013), Liu et al. (2005), and Rose (2016; 2020). These applications take a word list as input and construct MCC question items to test knowledge of those specific words relative to each other.

Well-Formedness

No matter whether MCC question items are created manually or using an automated generation system, the aim is the same: to produce well-formed MCC questions. Condensing from the various works described above, the following conditions describe the basic conditions we assume for well-formed MCC questions:

A. The stem sentence is not more difficult than the level of the key.
B. The stem sentence is not too difficult for the target test-takers.
C. The key is being used in a syntactic context and semantic sense that is common to the word (unless testing obscure uses of the word is desired).
D. The distractors are clearly sub-optimal choices for linguistic reasons, relative to the key.

Conditions A and B are necessary to ensure that the learner's knowledge of the key is the sole focus of the assessment and not additionally their ability to comprehend the stem sentence itself (cf., Jiang & Lee, 2017). Conditions C and D both have a certain amount of intended vagueness in them which allow for adaptation of the difficulty level of items. That is, testing obscure uses of the word may be pedagogically desired at times (cf., Ehara, 2022). Furthermore, testing whether learners can distinguish between the key and a slightly sub-optimal distractor may also be pedagogically desired (cf., Ondov et al., 2024).

Research Question

While the various systems described above show great potential and reported validation studies show that they can produce usable items, various limitations exist with these tools. For instance, for all the tools, post-generation hand-checking is still

required, lengthy generation time might still be required (even if automated), the tools themselves may not be easily available, and even if available, they may be difficult for non-programmers to use.

Recent advances in large language models (LLMs) such as OpenAI's GPT models and Google AI's Gemini models provide excellent tools for text generation and can help to overcome many of the limitations described above with previous systems (e.g., see Scaria et al., 2024 for a field-specific question generation system and Biancini et al., 2024 for a multiple-choice question generation system from source texts). The present work seeks to leverage the use of LLMs to generate well-formed MCC questions for use in second language vocabulary knowledge assessment. The generator tool, VocQGen will be described in detail in the next section and then results of a validation test which compares VocQGen-produced MCC questions to those produced manually. Hence, the fundamental research question of the present work is as follows:

RQ: How do VocQGen-produced MCC question items compare to manually-produced MCC question items?

Introduction to VocQGen

VocQGen (first described in Wang et al., 2023) was developed using the Python programming language (version 3.11; https://www.python.org/downloads/release/python-3119/). It comprises three pipelines: the "pre-processing pipeline" for creating a word group for each input word, including its commonly used part-of-speech (POS) tags and inflected forms; the "stem pipeline" for generating and validating question stems; and the "distractor pipeline" for selecting and validating distractors from the same vocabulary unit. The source code of the system can be found on Github at: https://github.com/judywq/VocQGen.

Pre-Processing Pipeline

In this pipeline, two POS tagging libraries, LemmInflect (https://github.com/bjascob/LemmInflect) and UniMorph (https://github.com/unimorph) are utilized together with Google Ngrams API (https://books.google.com/ngrams/). The use of these two POS tagging libraries ensures a more comprehensive and accurate identification of POS tags. Google Ngrams provides frequency data on the usage of a word under specific POS tags, allowing for the exclusion of tags that are infrequently used. For each input word, the two libraries assign all applicable POS tags, which are combined to form a "union set" of POS tags. This set is then ranked by frequency using data from Google Ngrams. If a POS tag ranked N has a frequency less than 10% of the next higher tag ranked N–1, it and any lower-ranked tags are removed from the set, leaving only the most common and relevant POS tags.

Next, for each retained POS tag, the corresponding inflected forms are incorporated. For example, nouns are expanded to include their plural forms (e.g., NNS), and verbs are expanded to include various tenses and non-finite forms (e.g., VBZ for third-person singular present, VBG for present participle). This process results in the creation of a

"word group" that encompasses the central word along with its most common POS tags and inflected forms.

Stem Pipeline

In generating the stem, the first step is to select a "key", i.e., the correct answer to the question to be generated. This is done by randomly selecting a POS tag and an inflected form of a word from its word group, with the possibility of selecting the base form (indicated by zero inflection). For example, for the word "account" with POS tags including NN (noun) and VB (verb), the NN tag may firstly be randomly chosen, and subsequently between the singular and plural form of the noun, the plural form "accounts" (NNS) may be selected as the key. After that, the information (Key: "accounts", inflection-included POS: "NNS") is fed forward for stem generation.

Once the key is determined, GPT-4, the base model for the popular application ChatGPT-4 (https://openai.com/index/gpt-4/) is prompted to generate a sentence using the selected POS and inflected form. After the sentence is generated, Python scripts validate the sentence mechanics—such as spelling and the position of the key—while spaCy (https://spacy.io/), an NLP tool, verifies the POS of each word in the sentence. If the sentence passes validation, the key is replaced with a blank, finalizing the stem. If the sentence fails validation, the process returns to the selection of a new key.

Distractor Pipeline

In the distractor pipeline, ten words sharing the same POS as the key are randomly selected from word groups other than that of the key, within the same user-specified input word list. GPT-4 then assesses each candidate for syntactic compatibility in the blank (to ensure grammatical correctness) and semantic appropriateness (distractors should make less sense than the key, making them sub-optimal choices). Words that are syntactically compatible but semantically sub-optimal are marked as valid distractors.

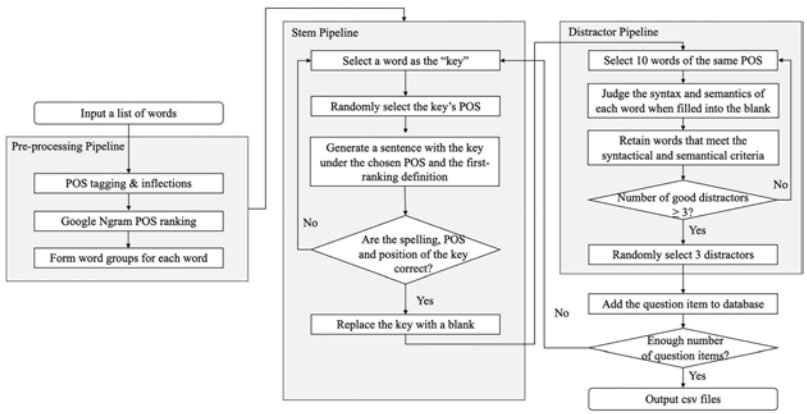

Figure 1 *VocQGen Flowchart.*

The program requires a minimum of three valid distractors; if this threshold is met, three are randomly chosen for the final question. If fewer than three valid distractors are identified, the validation process repeats. The overall generation process of VocQGen is summarized in Figure 1.

Prompting Techniques

In stem generation and distractor selection, we implemented few-shot prompting by providing GPT-4 with an example for each task to enhance its performance in these areas. For stem generation, we introduced specific variables that allow users to tailor to their unique contexts. Table 1 outlines these variables and their respective definitions. In distractor selection, we defined two primary variables: "sentence", which denotes the stem generated by GPT-4, and "words_with_comma", which represents a list of comma-separated distractors.

Table 1 *Variables in Stem Generation*

Variable name	Meaning
Word	The key/answer based on which the stem should be generated
Tag	Inflection-included POS tag of the key
Domain	Domain of English, such as English for General Academic Purposes
Scope	The field of study for the stem sentences, such as biology
Level_start	The lowest proficiency of learners based on CEFR
Level_end	The highest proficiency of learners based on CEFR
Length	The length of the stem sentence, such as 15–20 words

```
You are an English teacher at a Japanese university and you are creating
exemplary sentences to show your students the use of specific English words
in the domain of {domain}. The English proficiency of your students ranges from
{level_start} to {level_end} based on CEFR.
```

```
Now please create a sentence that meets the following criteria:
The sentence should show the use of the word "{word}" tagged as "{tag}".
The word "{word}" should be pivotal to the meaning of the sentence and
carries significant weight in the context.
The sentence should show a common usage of the word "{word}" tagged as "{tag}".
The sentence may be contextualized in {scope}.
The sentence should be understandable to the target students,
with no words beyond their proficiency levels.
The length of the sentence should be {length}.
Ensure "{word}" does not appear at the beginning or the end of the sentence,
nor is it repeated elsewhere in the sentence.
Ensure none of the derivatives of "{word}" are present in the sentence.
Please avoid starting the sentence with the definite article "the"
as much as possible.
```

```
Example: If the provided word is "account", tagged as "NN",
an appropriate sentence can be:
In managing finances, maintaining a bank account is essential for every student.
```

Figure 2 *Prompt for Stem Generation.*

```
You are a university English teacher and you are creating distractors for
vocabulary multiple-choice cloze questions for your students.
In this question stem: "{sentence}"
A list of possible distractors includes "{words_with_comma}".
```

```
Please evaluate the syntactic appropriateness/grammatical accuracy and
contextual appropriateness/semantic sense-making of the distractors in
the completed sentences.
Return only the following result in JSON format:
{{
  "distractor 1": {{"syntax": true, "semantics": true}},
  "distractor 2": {{"syntax": true, "semantics": false}}
}}
```

```
Example: In the question stem: "Birds ___ in the sky." the key is "fly".
The list of distractors includes "swim, pick". The distractor "swim"
is syntactically appropriate because there will be no grammar errors
when it is filled into the blank, but it does not make sense since birds
"fly" in the sky, not "swim". The distractor "pick" is syntactically
inappropriate because "pick" is a transitive verb and requires an
object after it. There will be  grammar errors when it is filled into
the blank. It also does not make much sense. Thus the returned result
in JSON format is as follows:
{{
  "swim": {{"syntax": true, "semantics": false}},
  "pick": {{"syntax": false, "semantics": false}}
}}
```

Figure 3 *Prompt for Distractor Selection.*

The prompt for stem generation is illustrated in Figure 2, while the prompt for distractor validation is depicted in Figure 3. Both prompts are structured into three segments: a background section (highlighted in blue) providing the context, a request section (in orange) articulating the task, and an example section (in green) serving as the example, or "shot", for few-shot prompting, and illustrating the desired response format expected from GPT-4 with two sample distractors.

Methods

To address our research question, we generated MCC items using VocQGen and collected manually created questions for comparison. The comparison focused on the well-formedness, i.e., appropriateness, of question stems and distractors, as well as the response patterns of potential test-takers. Expert reviewers were engaged to evaluate the well-formedness of the items, and a pilot testing experiment was conducted with students to examine the response patterns. The research protocol was approved by the Waseda University internal review board.

Question Creation

For the VocQGen items, we generated questions based on words from the first sublist of the Academic Word List (AWL; Coxhead, 2000) as it is a legacy from our previous work (Wang et al., 2023), though other word lists could be used (e.g., the New

Academic Word List; Browne et al., 2013). The AWL is frequently used in academic English courses for English as a Second Language (ESL) learners at tertiary institutions and comprises 570 word families organized into 10 sublists. The list contains only content words, including nouns, verbs, adjectives and adverbs, while functional words such as articles, pronouns and prepositions are not present. Sublist 1 includes 60 headwords along with their derivatives. We randomly selected one word from each family and generated one question for a randomly selected POS and inflected form of the word.

In configuring the prompts for stem generation, we tailored the variables to align with the proficiency levels of students at a Japanese university who were the potential test-takers, ranging from A1 to B2. The focus was on English for General Academic Purposes, and the designated length for each stem was set at 15–20 words.

For the manually created items, we sourced questions from two teachers from the same university who were teaching courses where AWL words were part of the course goals and had previously created their own questions for testing their students. In the course they were teaching, each vocabulary unit test consisted of 20 questions. We gathered one unit from each teacher, resulting in 40 manual questions in total. Then we combined these 40 manually created items with the 60 automatically generated items, assigned each item an ID, and shuffled them for subsequent experimentation and analysis.

Expert Review

The 100 questions underwent a review process conducted by two experts, both native English speakers with over 30 years of experience teaching English at the tertiary level in Japan. Initially, the experts assessed the grammatical accuracy of the stems and whether they accurately reflected common usage of the keys. Upon passing this evaluation, the experts further verified whether the POS of the distractors matched those of the keys and whether any distractor was a better option than the key. Comments for non-well-formed stems/distractors were encouraged for later qualitative analysis. In cases where the two experts disagreed, a third expert was consulted to provide a final judgment.

Quantitatively, we calculated the inter-rater agreement between the two experts using Cohen's Kappa coefficient as the primary metric, supplemented by percentage agreement. After the third expert resolved any disagreements, we separated the questions into manually generated and autogenerated sets and compared the final review results. Beyond the quantitative analysis, a thematic analysis was conducted on the comments provided by the experts. The researchers adopted an inductive coding approach to identify and categorize any salient issues or patterns within the questions.

Pilot Testing Experiment

For the testing experiment, we conducted an online assessment using Microsoft Forms with 68 student participants from Waseda University. Informed consent was obtained from all participants prior to the experiment.

Using the students' responses, we first calculated the number of correct responses for each question. Then, we conducted a statistical significance test to compare correct response rates between questions in the two sets. Finally, we analyzed some of the least and most challenging items to understand the factors contributing to their high or low correct response rates.

Results

Expert Review

The evaluation of question stems revealed a 93% agreement and a Cohen's Kappa $\kappa = 0.50$ between the two experts. For distractors, the percentage agreement was 91.95% while the Cohen's Kappa $\kappa = 0.51$. Following the resolution of disagreements, the number of valid and invalid stems and the number of appropriate (well-formed) and inappropriate (non-well-formed) distractors (excluding distractors for stems deemed inappropriate) for both manually and automatically generated items are summarized in Table 2.

The results show that automatically generated question items had a higher proportion of appropriate stems (95.00%) and distractors (95.32%) compared to manually generated items (85.00% appropriate stems and 80.39% appropriate distractors). In the following, examples of inappropriate stems and distractors will be presented for qualitative analysis.

Stems

For the automatically generated questions, three stems were identified as inappropriate due to issues with semantics. Two of these stems exhibited awkward usage of the modal verb "must," and one displayed poor logical flow. The following shows all three invalid stems.

- ***Stem**: "Teachers must fairly ____ textbooks to all students on the first day of class."*
  ***Key**: distribute*
- ***Stem**: "Students must often collaborate to study various subjects and improve skills in a specific ____ of interest."*
  ***Key**: area*
- ***Stem**: "Water is composed of two hydrogen atoms bonded to one oxygen atom, making it a simple yet vital ____ of life."*
  ***Key:** constituent*

Table 2 *Number of Appropriate and Inappropriate Stems and Distractors*

	Stem			Distractor		
	Appropriate	Inappropriate	%	Appropriate	Inappropriate	%
Manual	34	6	85.00%	82	20	80.39%
Auto	57	3	95.00%	163	8	95.32%

In the first and second examples, the use of "must" is awkward because it suggests a stronger obligation than what is contextually appropriate. In the third example, the logical flow is problematic, as it implies that the composition of water is directly related to its significance in life, which is not a natural reasoning process.

For manually generated questions, most inappropriate stems were related to insufficient context, leading to confusion or diminishing the significance of the key in the stem. The following are two typical examples out of the 6 inappropriate items.

- ***Stem**: "All over the railway system, _____ expensive bridges and tunnels are nearing collapse."*
  ***Key:** similarly*
- ***Stem**: "On the one hand, they are a concentrated _____ of energy."*
  ***Key:** source*

In the first example, the stem is confusing due to the vague contextual link between "similarly" and the rest of the sentence. In the second example, the use of "they" is decontextualized, making the key "source" less intuitive and less anticipated.

Distractors

For the distractors in the automatically generated questions, two were deemed inappropriate due to grammar errors, specifically related to bare nouns (singular nouns not preceded by any determiner). The remaining six inappropriate distractors were found to be acceptable answers, which compromised the effectiveness of the questions. The following shows one example each.

- ***Stem**: "Many companies have ____ their operations to improve efficiency and productivity." –*
  ***Key:** restructured; **Inappropriate distractor:** required*
- ***Stem**: "To achieve consistent results in an experiment, minimizing ____ is crucial for researchers."*
  ***Key**: variance; **Inappropriate distractor**: indicator*

In the first example, "required" is an acceptable answer given the common collocation of "require sb. to do sth." Besides, "required" also makes sense when filled into the blank. In the second example, English grammar requires singular countable nouns to be preceded by determiners except for in fixed collocations. Here, "indicator" is a singular noun not preceded by any determiner, which results in a grammar error when the distractor is filled into the blank. The key, "variance" with the POS tag of NN (a noun in its singular form), however, is not affected as it is an uncountable noun. This points to a lack of consideration in the system regarding countable and uncountable nouns.

For the manually generated questions, 19 distractors were inappropriate due to grammar errors, 16 of which involved unfit POS of the distractors. Only one distractor was considered an acceptable alternative answer. The following is an example of a distractor of unfit POS:

- ***Stem**: "Children need something to _____ their time; otherwise, they get bored."*
  ***Key**: occupy; **Inappropriate distractor**: status*

In this example, the distractor "status" is a noun, whereas the correct answer "occupy" is a verb. The grammatical mismatch could lead students to dismiss the distractor without engaging with the question meaningfully, undermining the validity of the item.

Testing Experiment

Figure 4 illustrates the distribution of correct responses for the two test sets. The boxplot on the left shows the number of correct responses, while the violin plot on the right visualized the density of the number of correct responses.

The results indicate that automatically generated questions generally had more correct responses, as evidenced by the medians in the boxplot and the clustering of questions at the higher end of the violin plot. In contrast, the number of correct responses for manually generated questions exhibits a wider vertical spread, as depicted in the violin plot, suggesting greater variability among individual questions and more balanced distribution.

A Shapiro-Wilk test was conducted to assess the normality of the correct response distributions for both sets of questions. The test revealed that the distribution was not normal for either the auto-generated items ($W = 0.903$, $p < 0.001$) or the manually generated items ($W = 0.927$, $p = 0.01$). Thus, a Mann–Whitney U-test was employed to compare the two test sets. The results demonstrated a statistically significant difference between the two sets ($U = 1692.0$, $p = 0.00$), indicating that the auto-generated and manually generated questions differ significantly in terms of their correct response rates.

To further explore this difference, we analyzed the questions with the lowest and highest correct response rates from both sets. For the questions with the highest correct response rates, we found that both sets had keys and distractors that were clearly distinct in meaning, making the distractors easy to dismiss. Particularly in the manually generated set, distractors could also frequently be rejected based on grammatical grounds. The following presents the question with the highest rates in each set.

Autogenerated (with 68 correct responses):

- ***Stem****: "Stress is a significant ____ affecting students' academic performance and well-being."*
  Key*: factor;* ***Distractors****: estimation, section, individualist*

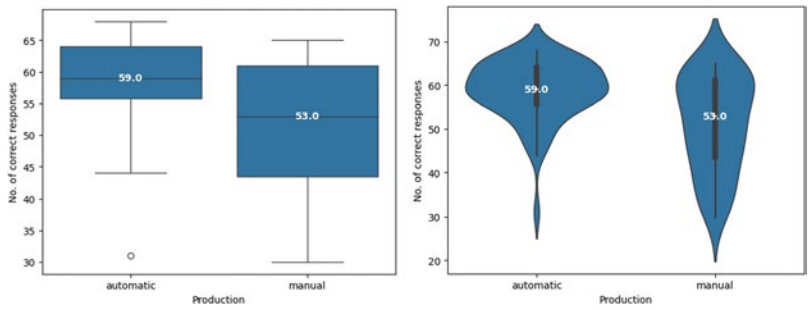

Figure 4 *Number of Correct Responses for the Two Test Sets.*

Manually generated (with 65 correct responses):

- ***Stem**: "Doctors and lawyers are high _____ jobs in our society."*
  ***Key**: status; **Distractors**: access, promote, internal*

For the questions with the lowest correct response rates in the autogenerated set, distractors that were deemed acceptable alternatives by the experts were the major reason. In contrast, in the manually generated set, it was often due to a lack of context or the use of a key in an unfamiliar context for the test takers. The following presents the questions of the lowest rates from each set.

Autogenerated (only 31 correct responses):

- ***Stem**: "She missed the lecture due to a cold and will ____ need to catch up with her classmates."*
  ***Key**: evidently; **Distractors**: environmentally, constitutionally, consistently*

Manually generated (only 30 correct responses):

- ***Stem**: "Reference books may not be _____ from the library."*
  ***Key**: removed; **Distractors**: sequenced, movable, publishable*

In the first example, the distractor "consistently" is deemed an acceptable answer by experts and was indeed chosen by many test takers. The seemingly challenging question has a flaw in question design. In the second example, "removed" is the most suitable answer but test takers may not be familiar with "reference books" in the context.

Discussion

The results of this study indicate that automatically generated questions exhibit higher proportions of appropriate question stems and distractors compared to manually generated ones. Specifically, the stems of the automatically generated questions are more contextualized and free from grammatical errors. However, some issues were noted in the semantics and logical coherence of the stems. For distractors, while the automatically generated questions performed well in many respects, there were instances where distractors were deemed acceptable answers, a problem less frequently observed in manually generated items. Table 3 summarizes the pros and cons of automatically generated and manually generated items.

The main strength of automatically generated questions lies in the accuracy and contextual relevance of the stems. The use of large language models (LLMs) within

Table 3 *Strengths and Weaknesses of Question Generation Approaches*

Approach	Pros	Cons
Automatic	Stems are free from grammar errors and are contextualized	Some distractors are acceptable answers
Manual	Distractors are rarely acceptable answers	Some stems are decontextualized and confusing; some distractors do not fit grammatically into the blanks

the system ensures that grammatical correctness is maintained and that the stems are contextually appropriate. However, LLMs currently lack the nuanced judgment required to ensure that distractors are semantically unfit as answers. This limitation stems from the fact that current LLMs, while capable of processing language with remarkable accuracy, still fall short of replicating human-like reasoning in determining the subtle appropriateness of distractors. Unlike POS validation, which can be performed using rule-based NLP tools, assessing the semantic appropriateness of distractors still requires human reasoning. Therefore, human intervention remains crucial in refining automatically generated questions, particularly in reviewing distractors.

As for the lack of challenge in the autogenerated questions, two primary factors may have contributed to this. First, the clarity of context in the autogenerated questions may have made it easier for students to understand and answer correctly. Second, and more importantly, the distractors were often too easily ruled out. This issue arises from the inherent design of the system, which generates distractors from the same vocabulary unit or sublist as the key. Given the limited pool of words within these units, finding distractors that are both similar to the correct answer and semantically suboptimal can be challenging. Consequently, students who have a basic understanding of the key's meaning can often easily identify the correct answer, without having to distinguish the nuances between a key and distractors. Given this limitation, the autogenerated questions may serve better in practice quizzes, i.e., no-stakes tests, rather than high-stakes tests expected to set students of different proficiencies apart.

Limitations

This study's findings are subject to several limitations. First, the manually generated questions were collected from previous teachers, and the amount of effort those teachers invested in crafting these questions is unknown. Some of the manually generated questions may have been hastily created or sourced from online texts without thorough proofreading. It is possible that with more time and careful effort, the proportion of appropriate questions could surpass that of automatically generated ones. The small number of teachers might have also introduced personal biases into the questions and subsequently undermined the validity of the research results.

Second, the evaluation study was conducted using only one sublist of the AWL words. The results may vary when different sublists or other word lists are used. The AWL is unique in that it is not an elementary list like the General Service List (GSL; West, 1953). All words in the AWL are content words, limited to nouns, verbs, adjectives, and adverbs, while function words such as pronouns and articles are absent. For more elementary word lists like the GSL, creating stems that highlight function words and finding appropriate distractors of the same POS may present additional challenges, particularly with function words like articles.

Finally, the vocabulary proficiency levels of the students were not assessed before the experiment. As a result, it is not possible to evaluate how the manual and autogenerated questions correlate with and thus reflect students' proficiency levels. Future studies should consider including a pre-assessment of students' proficiency levels to better understand the applicability and effectiveness of both automatically and manually generated vocabulary questions across different proficiency levels.

Conclusion

This study explored the effectiveness of an automated vocabulary quiz generation tool, VocQGen, compared to manually created questions. The findings revealed that automatically generated questions generally have higher quality, particularly in terms of grammatical accuracy and contextual relevance of stems. However, challenges remain in ensuring that distractors are semantically unfit and do not serve as acceptable answers, an area where manually generated questions tend to perform better. In addition, automatically generated questions are significantly less challenging than manually generated ones, with the number of correct responses for individual questions cluster at the higher spectrum. Future work should focus on improving the system's ability to generate more effective distractors and exploring methods to make the questions more challenging. Despite these limitations, the automated system offers significant advantages in terms of time efficiency and consistency. Nevertheless, human review of such questions is still needed before they can be used for testing purposes.

Acknowledgments

This paper is a part of research outcomes funded by Waseda University Grant for Special Research Project (project number: 2023R-020). We also gratefully acknowledge the cooperation of the two raters in this work as well as the 68 students who participated in the study.

References

Abdullah, M., Madain, A., & Jararweh, Y. (2022). ChatGPT: Fundamentals, applications and social impacts. *2022 Ninth International Conference on Social Networks Analysis, Management and Security (SNAMS),* 1–8. https://doi.org/10.1109/SNAMS58071.2022.10062688

Abu-Alhija, F. N. (2007). Large-scale testing: Benefits and pitfalls. *Studies in Educational Evaluation,* 33(1), 50–68. https://doi.org/10.1016/j.stueduc.2007.01.005

Alqahtani, M. (2015). The importance of vocabulary in language learning and how to be taught. *International Journal of Teaching and Education,* 3(3), 21–34. https://doi.org/10.52950/TE.2015.3.3.002

Biancini, G., Ferrato, A., & Limongelli, C. (2024). Multiple-choice question generation using large language models: Methodology and educator insights. *Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization,* 584–590. Presented at the Cagliari, Italy. https://doi.org/10.1145/3631700.3665233

Brown, J., Frishkoff, G., & Eskenazi, M. (2005). Automatic question generation for vocabulary assessment. *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing,* 819–826. https://aclanthology.org/H05-1103/

Brown, T. S., & Perry, F. L. (1991). A comparison of three learning strategies for ESL vocabulary acquisition. *TESOL Quarterly,* 25(4), 655–670. https://doi.org/10.2307/3587081

Browne, C., Culligan, B. and Phillips, J. (2013). *The new academic word list*. http://www.newgeneralservicelist.org/nawl-new-academic-word-list/

Coniam, D. (1997). A computerised English language proofing cloze program. *Computer Assisted Language Learning,* 10(1), 83–97. https://doi.org/10.1080/0958822970100106

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213–238. https://doi.org/10.2307/3587951

Ehara, Y. (2022). No meaning left unlearned: Predicting learners' knowledge of atypical meanings of words from vocabulary tests for their typical meanings. In A. Mitrovic & N. Bosch (Eds.), *Proceedings of the 15th International Conference on Educational Data Mining,* 492–499. https://eric.ed.gov/?id=ED624124

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. Routledge.

Goto, T., Kojiri, T., Watanabe, T., Iwata, T., & Yamada, T. (2010). Automatic generation system of multiple-choice cloze questions and its evaluation. *Knowledge Management & E-Learning: An International Journal,* 2(3), 210–224. https://www.kmel-journal.org/ojs/index.php/online-publication/article/view/78

Hale, G. A., Stansfield, C. W., Rock, D. A., Hicks, M. M., Butler, F. A., & John W. Oller, J. R. (1989). The relation of multiple-choice cloze items to the Test of English as a Foreign Language. *Language Testing,* 6(1), 47–76. https://doi.org/10.1177/026553228900600106

Jiang, S., & Lee, J. (2017, December). Carrier sentence selection for fill-in-the-blank items. In Y.-H. Tseng, H.-H. Chen, L.-H. Lee, & L.-C. Yu (Eds.), *Proceedings of the 4th Workshop on Natural Language Processing Techniques for Educational Applications* (NLPTEA 2017) (pp. 17–22). https://aclanthology.org/W17-5903

Khoii, R., & Sharififar, S. (01 2013). Memorization versus semantic mapping in L2 vocabulary acquisition. *ELT Journal,* 67(2), 199–209. https://doi.org/10.1093/elt/ccs101

Kunichika, H., Katayama, T., Hirashima, T., & 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),* 1117–1124. https://api.semanticscholar.org/CorpusID:1536529

Lee, K., Kweon, S., Kim, H., & Lee, G. (2013). Filtering-based automatic cloze test generation. *Proceedings of Speech and Language Technology in Education (SLaTE),* 72–76.

Liu, C., Wang, C., Gao, Z., & Huang, S. (2005). Applications of lexical information for algorithmically composing multiple-choice cloze items. *Proceedings of the 2nd Workshop on Building Educational Applications Using NLP,* 1–8. https://aclanthology.org/W05-0201/

Mitkov, R., Ha, L. A., & Karamanis, N. (2006). A computer-aided environment for generating multiple-choice test items. *Natural Language Engineering,* 12(2), 177–194. https://doi.org/10.1017/S1351324906004177

Mitkov, R., Ha, L. A., Varga, A., & Rello, L. (2009, March). Semantic similarity of distractors in multiple-choice tests: Extrinsic evaluation. In R. Basili & M. Pennacchiotti (Eds.), *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics* (pp. 49–56). https://aclanthology.org/W09-0207

Nation, P. (2022). Teaching and learning vocabulary. In E. Hinkel (Ed.), *Handbook of practical second language teaching and learning* (pp. 397–408). Routledge.

Ondov, B., Attal, K., & Demner-Fushman, D. (2024, June). Pedagogically aligned objectives create reliable automatic cloze tests. In K. Duh, H. Gomez, & S. Bethard (Eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 3961–3972). https://doi.org/10.18653/v1/2024.naacl-long.220

O'Reilly, R. P., & Streeter, R. E. (1977). Report on the development and validation of a system for measuring literal comprehension in a multiple-choice cloze format: Preliminary factor analytic results. *Journal of Reading Behavior,* 9(1), 45–69. https://eric.ed.gov/?id=ED129878

Pino, J., Heilman, M., & Eskenazi, M. (2008). A selection strategy to improve cloze question quality. *Proceedings of the Workshop on Intelligent Tutoring Systems for Ill-Defined Domains. 9th International Conference on Intelligent Tutoring Systems,* Montreal, Canada, 22–32.

Rose, R. (2016). Automatic word quiz construction using regular and simple English Wikipedia. *INTED2016 Proceedings,* 8032–8040. https://doi.org/10.21125/inted.2016.0889

Rose, R. (2020). Improving the production efficiency and well-formedness of automatically-generated multiple-choice cloze vocabulary questions. *Proceedings of the 12th Language Resources and Evaluation Conference,* 7094–7101. https://aclanthology.org/2020.lrec-1.877/

Rose, R., Sugawara, A., Orita, N., & Wang, Q. (2022). Evaluation dataset of multiple-choice cloze items for vocabulary training and testing. *Adjunct Proceedings of the 2022 ACM International Joint Conference on Pervasive and Ubiquitous Computing and the 2022 ACM International Symposium on Wearable Computers,* 292–298. https://doi.org/10.1145/3544793.3560378

Sagarra, N., & 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. https://doi.org/10.1111/j.1540-4781.2006.00394.x

Scaria, N., Dharani Chenna, S., & Subramani, D. (2024). Automated educational question generation at different Bloom's skill levels using large language models: Strategies and evaluation. In A. M. Olney, I.-A. Chounta, Z. Liu, O. C. Santos, & I. I. Bittencourt (Eds.), *Artificial Intelligence in Education* (pp. 165–179). Springer Nature Switzerland.

Schmitt, N. (1997). Vocabulary learning strategies. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 199–227). Cambridge University Press.

Sumita, E., Sugaya, F., & Yamamoto, S. (2005). Measuring non-native speakers' proficiency of English by using a test with automatically-generated fill-in-the-blank questions. *Proceedings of the Second Workshop on Building Educational Applications Using NLP,* 61–68. Presented at the Ann Arbor, Michigan. USA: Association for Computational Linguistics.

Taylor, W. L. (1953). "Cloze procedure": A new tool for measuring readability. *Journalism Quarterly,* 30(4), 415–433. https://doi.org/10.1177/107769905303000401

Wang, Q., Rose, R., Sugawara, A., & Orita, N. (2023). Automated generation of multiple-choice cloze questions for assessing English vocabulary using GPT-turbo 3.5. *The Joint 3rd International Conference on Natural Language Processing for Digital Humanities and 8th International Workshop on Computational Linguistics for Uralic Languages*, 52–61. https://aclanthology.org/2023.nlp4dh-1.7

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Palgrave-Macmillan.

West, M. (1953). *A general service list of English words*. Longman, Green and Co.

Zhang, Z., Mita, M., & Komachi, M. (2023, May). Cloze quality estimation for language assessment. In A. Vlachos & I. Augenstein (Eds.), *Findings of the Association for Computational Linguistics: EACL 2023* (pp. 540–550). https://doi.org/10.18653/v1/2023.findings-eacl.39