Generating Competitive Distractors from Student Error Data

Myrthe Braam
MemoryLab
myrthe@memorylab.nl

Maarten van der Velde
MemoryLab
maarten@memorylab.nl

Hedderik van Rijn
University of Groningen
hedderik@van-rijn.org

ABSTRACT

While students often prefer learning with multiple-choice questions (MCQs), these are frequently criticized for emphasizing recognition over active recall, which makes them less effective for long-term retention. However, well-designed competitive distractors, that are semantically or orthographically similar to the correct answer, may overcome this deficit by encouraging deeper cognitive processing. Traditional distractor generation is labor-intensive, requiring domain expertise and significant time investment. Recent advances in artificial intelligence offer automated solutions for generating semantically related distractors at scale, but not for generating distractors that target learners' misconceptions about the lexical representation of answers. In this study, we present a data-driven method for automatically generating plausible orthographically related distractors, based on a combination of common incorrect responses from learners doing open-answer retrieval practice, and rule-based generation of common misspellings. This flexible and scalable approach to generating distractors that align with genuine misconceptions, can make MCQs a more effective tool for vocabulary learning.

Keywords

distractor generation, multiple-choice questions, orthographically related distractors, retrieval practice

1. INTRODUCTION

Multiple-choice questions (MCQs) are commonly used in retrieval practice due to their efficiency, accessibility and ease of implementation. However, they are often criticized for primarily testing recognition rather than active recall, which is generally considered more effective for learning. Despite this skepticism, research suggests that well-constructed MCQs, particularly those with carefully designed distractors, can enhance learning outcomes by promoting deeper cognitive processing [16, 17].

A key concern with traditional MCQs is that they allow learners to recognize the correct answer rather than actively retrieve it, which can be less effective for long-term retention [4, 12]. This passive processing enables learners to guess answers without fully recalling the information. For example, in a vocabulary learning context, if there is only one answer option that looks vaguely similar to the cue, or if there is one noun combined with three adjectives, learners may guess it correctly without truly engaging with the material. This can result in learners overlooking crucial details like spelling, making these MCQs insufficient preparation for tests where accurate spelling is assessed [19].

However, MCQs offer several advantages from a practical standpoint as they are time-efficient and reduce cognitive load compared to open-ended questions, making them particularly useful in adaptive learning systems. Research suggests that students often avoid practice methods that demand greater mental effort, such as open-ended retrieval practice [9]. Therefore, when designed effectively, MCQs can serve as a more accessible and engaging retrieval practice tool, increasing the likelihood of student use.

Creating high quality MCQs, however, can be time consuming and requires significant expertise to ensure that distractors are neither too obvious nor entirely misleading. Furthermore, the relevance of distractor types depends on the specific learning goal, for instance, mastering spelling requires different distractors than those used for understanding the conceptual meaning of a word. In recent years, new techniques have been developed to automate distractor generation, offering promising solutions. However, reliably producing high quality, contextually appropriate distractors remain a difficult task. In this study, we present a structured approach to automatically generate plausible distractors for vocabulary retrieval practice, using common student mistakes.

2. LEARNING WITH MCQS

2.1 Good Distractors

Recent research suggests that competitive distractors, those closely related to the correct answer, can introduce desirable difficulty, forcing learners to engage in deeper processing [17, 2, 3]. Effective distractors are plausible yet incorrect, requiring the learner to critically evaluate their choices. For instance, the use of semantic similarity (using words with related meanings) or orthographic similarity (using words with similar spelling patterns) encourages learners to engage in retrieval processes rather than simple elimination, thereby strengthening their understanding of the material [16, 17]. This makes MCQs more comparable to open-ended questions in terms of their learning benefits.

Importantly, the effectiveness of distractors may depend on learners' proficiency levels and learning goals. Research [15] suggests that sensitivity to confusion caused by similar lexical forms decreases as language proficiency increases. In other words, lower proficiency learners are more susceptible to confusion from orthographically similar distractors, while higher proficiency learners may benefit from them. Therefore, aligning distractor design with learners' proficiency levels and learning objectives is essential. For instance, if the goal is to enhance semantic understanding, distractors focusing on meaning may be more

Myrthe Braam, Maarten van der Velde, and Hedderik van Rijn. Generating Competitive Distractors from Student Error Data. In Caitlin Mills, Giora Alexandron, Davide Taibi, Giosuè Lo Bosco, and Luc Paquette (eds.) Proceedings of the 18th International Conference on Educational Data Mining, Palermo, Italy, July, 2025, pp. 592–595. International Educational Data Mining Society (2025).

https://doi.org/10.5281/zenodo.15870234

effective. In contrast, orthographically related distractors can be more useful for reinforcing spelling accuracy, which often becomes a priority in later stages of learning.

Furthermore, studies have highlighted the importance of using students' misconceptions and common errors to create plausible distractors in multiple-choice testing. Traditionally, this process involved manually analyzing students' written or verbal responses to identify patterns of misunderstanding [13, 18]. Content specialists could then select the most frequently occurring mistakes from students' responses, or generate individual distractors by applying these common errors and misconceptions [6, 13].

Utilizing frequently made mistakes by learners has proven to be a useful technique for distractor generation [10, 23]. Nonetheless, some researchers have expressed concerns regarding the use of incorrectly spelled distractors, arguing that exposure to incorrect information may result in the acquisition of false knowledge [7, 14, 24].

Despite these concerns, other studies indicate that the negative effects of misleading distractors can be mitigated. It was found that providing students with immediate feedback effectively counteracted the negative impact of related distractors, enhancing learning outcomes [8]. Another study [2] further demonstrated the benefits of orthographically related distractors in vocabulary learning. Their study, which incorporated pseudowords and similar incorrect options in multiple-choice tests, showed that vocabulary recall improved both immediately and after a delay.

2.2 Distractor Generation

Traditionally, distractor generation has relied on manually creating or selecting incorrect answers focusing on certain key criteria: (1) Plausibility: the distractors should be similar to the correct answer, grammatically correct, and contextually relevant; (2) Reliability: each distractor should definitively be an incorrect answer; and (3) Diversity: the distractors should differ sufficiently from one another [1]. Creating such high-quality distractors is labor-intensive, requiring both domain expertise and significant time investment.

Recent advancements have introduced AI-based approaches to streamline this process. Pre-trained language models (PLMs) such as GPT, BERT, T5, and BART, trained on vast amounts of unlabeled data, can be adapted for distractor generation tasks [1]. These models can be designed to rank distractors based on their relevance to the question stem and correct answer or generate plausible distractors by considering the context of the material [20]. However, because these models primarily focus on semantic relationships and contextual meaning, they may struggle to produce orthographic distractors, which rely on surface-level similarities such as spelling variations.

Despite these advancements, AI-generated distractors still face challenges in ensuring reliability, i.e., whether the distractors are actually incorrect, and diversity, i.e., whether they differ sufficiently from each other [1]. Another significant limitation is that these AI-generated distractors may not accurately reflect the actual mistakes learners make during retrieval practice of vocabulary in a new language. While context and relatedness are considered, there is no direct connection between the errors learners commonly make and the distractors generated by these models, as they are generally trained on text corpora rather than on learners' errors. As a result, these methods are unlikely to fully capture learners' specific misconceptions about vocabulary materials. As noted earlier, distractors that accurately represent learners' misconceptions can promote more meaningful engagement and strengthen comprehension. Yet, identifying these misconceptions demands the analysis of real learner responses, which can be a complex and time-consuming process.

3. METHOD

To address these challenges, we developed an automatic distractor generation method for vocabulary learning. This method utilizes incorrect responses from learners to generate plausible orthographic distractors for vocabulary retrieval practice. The entire process is automated through a custom script written in R [21], using several key packages including 'data.table' [11] for efficient data manipulation, 'dplyr' [25] for data wrangling, and 'stringdist' [26] for string matching. The script is applied to a dataset of student responses collected from vocabulary learning sessions within an adaptive learning system. Importantly, this method is adaptable to any vocabulary retrieval practice setting, as long as learners' typed responses to retrieval cues are available.

3.1 Preprocessing stage

Before applying the steps for distractor generation, the dataset undergoes preprocessing. This involves retaining only the incorrect responses, removing empty and incomplete responses, and converting all text to lowercase.

3.2 Automatic Distractor Generation Steps

Our method follows these key steps (figure 1):

1) Filtering learning data on word overlap: We identify incorrect responses with a Damerau-Levenshtein similarity of ≥ 0.70 to the correct answer to ensure orthographic similarity.
2) Excluding nonexistent trigrams: To filter out typographical errors, we exclude the incorrect responses containing trigrams that do not exist in the second language. That way we distinguish between typing errors (mistakes caused by key confusion while typing) and spelling errors (cognitive errors due to insufficient language competence) [22].
3) Selecting the most common mistakes: High frequency incorrect responses are prioritized as distractors, as they reflect the most common learner misconceptions. For each correct answer, the corresponding incorrect responses are ranked based on their frequency.
4) Generating misspelled distractors: When all incorrect responses for a specific item are unsuitable for distractor selection, a misspelling generation function is applied. This function simulates real learner mistakes:
    a. Common consonant/vowel confusions (e.g., 'ie' vs. 'ei');
    b. Gender swaps (e.g., 'der' vs. 'die');
    c. Orthographic errors based on phonetic similarities (e.g., 'au' vs 'eu').
5) Select the distractors: Pick the highest frequency incorrect responses and/or the distractors generated by the misspelling function.

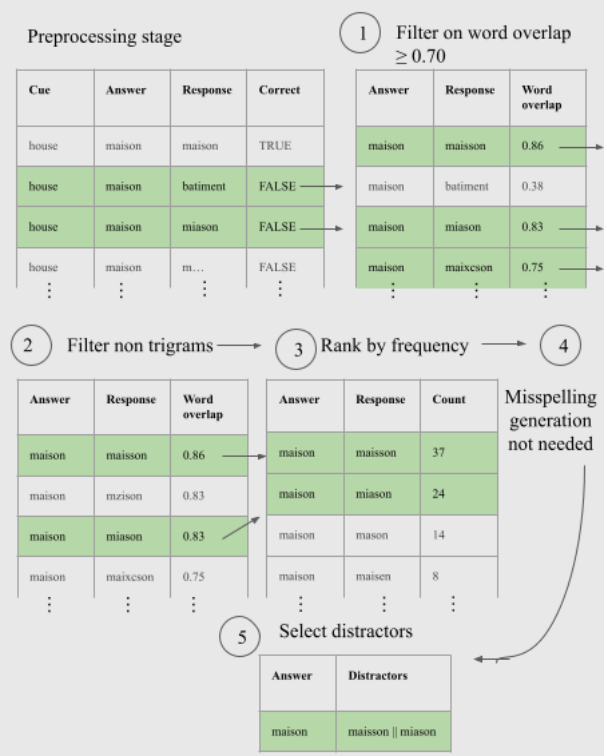

Figure 1. A visualization of the distractor generation method derived from response data.

4. RESULTS

We will present results from applying our distractor generation method to a large dataset of typed learner responses collected during German vocabulary practice in Dutch secondary schools. The dataset includes over 12 million responses from students across all grade levels and three educational tracks (preparatory secondary vocational education, senior general secondary education, and university preparatory education).

We focused on incorrect responses, which were cleaned and filtered to ensure they were plausible. Incomplete or empty entries were removed, and only those reasonably similar to the correct answer were retained. We also excluded responses with unlikely letter combinations for German. This resulted in a final set of over 117,000 plausible incorrect responses.

Using the procedure described in our method section, we grouped errors by vocabulary item and selected the most frequent ones as distractors. For 2,059 items, we aimed to select two distractors based on real learner data; where this was not possible, additional distractors were generated using a misspelling function. Of the 4,118 distractors selected, only 2.38% were generated.

We will present summary statistics and visualizations showing error frequencies, similarity to target words, and the extent to which distractors could be drawn from actual learner input.

5. DISCUSSION AND FUTURE DIRECTIONS

An essential next step is to evaluate the quality of the generated distractors. This can be achieved within real learning tasks by focusing on two key aspects. 1. Distractor selection rates, which measure how frequently learners choose a distractor as their answer. A high selection rate would suggest that the distractors are plausible and do not reveal the correct answer. 2. Learning outcomes, where enhanced long-term retention is expected, aligning with the concept of desirable difficulties in learning [17].

One line of work that we have not explored in this study is the extent to which this method generalizes to other populations than the population that generated the data. While both French and German learners of English need to learn the same orthographic representations, their native languages may influence the types of errors they make. For instance, it is likely that a French native speaker has different issues with a word such as *league* (from the French *ligue*, meaning 'confederacy') than a German speaker will have. The extent to which this plays a role is still to be determined.

Another important direction for development is the adaptation of distractors to learners' individual knowledge levels about the target language. Previous research suggests that an effective learning sequence in vocabulary learning involves first targeting semantic learning before transitioning to orthographic learning [5, 15].

6. CONCLUSION

Our automatic distractor generation method uses real learner mistakes to create plausible, competitive orthographic distractors for MCQs, with the aim of enhancing their effectiveness as a learning tool within adaptive vocabulary learning systems. This automated approach not only reduces the need for manual effort but also aims to generate high quality distractors that promote deeper cognitive processing and support more effective learning.

7. REFERENCES

[1] Alhazmi, E., Sheng, Q. Z., Zhang, W. E., Zaib, M., & Alhazmi, A. 2024. Distractor generation in multiple-choice tasks: A survey of methods, datasets, and evaluation. *arXiv*. DOI= https://doi.org/10.48550/ARXIV.2402.01512.

[2] Baxter, P., Bekkering, H., Dijkstra, T., Droop, M., Van den Hurk, M., & Leoné, F. 2022. Contrasting orthographically similar words facilitates adult second language vocabulary learning. *Learn. Instr*. 80, 101582. DOI= https://doi.org/10.1016/j.learninstruc.2022.101582.

[3] Baxter, P., Droop, M., Van Den Hurk, M., Bekkering, H., Dijkstra, T., & Leoné, F. 2021. Contrasting similar words facilitates second language vocabulary learning in children by sharpening lexical representations. *Front. Psychol.* 12, 688160. DOI= https://doi.org/10.3389/fpsyg.2021.688160.

[4] Bjork, R. A., & Whitten, W. B. 1974. Recency-sensitive retrieval processes in long-term free recall. *Cogn. Psychol.* 6(2), 173–189. DOI= https://doi.org/10.1016/0010-0285(74)90009-7.

[5] Braam, M., Wilschut, T., van der Velde, M., & van Rijn, H. 2024. Studying with optimized multiple-choice distractors equates recall-based studying. *eScholarship, University of California*. DOI= https://escholarship.org/uc/item/3j67575c.

[6] Briggs, D. C., Alonzo, A. C., Schwab, C., & Wilson, M. 2006. Diagnostic assessment with ordered multiple-choice items. *Educ. Assess.* 11, 1 (2006), 33–63. DOI= https://doi.org/10.1207/s15326977ea1101_2.

[7] Brown, A. S. 1988. Experiencing misspellings and spelling performance: Why wrong isn't right. *J. Educ. Psychol.* 80, (1988), 488–494.

[8] Butler, A. C., & Roediger, H. L. 2008. Feedback enhances the positive effects and reduces the negative effects of

multiple-choice testing. *Mem. Cogn.* 36, 3 (2008), 604–616. DOI= https://doi.org/10.3758/MC.36.3.604.

[9] Carpenter, S.K. 2023. Encouraging students to use retrieval practice: A review of emerging research from five types of interventions. *Educ.Psychol. Rev.* 35, 96. DOI= https://doi.org/10.1007/s10648-023-09811-8.

[10] De La Torre, J. 2009. A cognitive diagnosis model for cognitively based multiple-choice options. *Appl. Psychol. Meas*. 33, 3 (2009), 163–183. DOI= https://doi.org/10.1177/0146621608320523.

[11] Dowle, M., & Srinivasan, A. 2024. data.table: Extension of 'data.frame' *(R package version 1.16.4)* [Computer software]. DOI= https://doi.org/10.5281/zenodo.1234567.

[12] Glover, J. A. 1989. The "testing" phenomenon: Not gone but nearly forgotten. *J. Educ. Psychol.* 81(3), 392–399. DOI= https://doi.org/10.1037/0022-0663.81.3.392.

[13] Haladyna, T. M., & Rodriguez, M. C. 2013. Developing and Validating Test Items (0 ed.). *Routledge*. DOI= https://doi.org/10.4324/9780203850381.

[14] Jacoby, L. L., & Hollingshead, A. 1990. Reading student essays may be hazardous to your spelling: Effects of reading incorrectly and correctly spelled words. *Can. J. Psychol.* 44, 3 (1990), 345–358. DOI= https://doi.org/10.1037/h0084259.

[15] Kocić Stanković, A. 2008. The problem of synforms (similar lexical forms). *Facta Univ. Linguist. Lit.*, 6(1), 51-59.

[16] Little, J. L., Bjork, E. L., Bjork, R. A., & Angello, G. 2012. Multiple-choice tests exonerated, at least of some charges: Fostering test-induced learning and avoiding test-induced forgetting. *Psychol. Sci.* 23(11), 1337–1344. DOI= https://doi.org/10.1177/0956797612443370.

[17] Little, J. L., & Bjork, E. L. 2015. Optimizing multiple-choice tests as tools for learning. *Mem. Cogn*. 43(1), 14–26. DOI= https://doi.org/10.3758/s13421-014-0452-8.

[18] Moreno, R., Martínez, R. J., & Muñiz, J. 2015. Guidelines based on validity criteria for the development of multiple-choice items. *Psicothema* 27, (2015), 388–394. DOI= https://doi.org/10.7334/psicothema2015.110.

[19] Nakata, T. 2016. Effects of retrieval formats on second language vocabulary learning. *Int. Rev. Appl. Linguist. Lang. Teach.* 54, 3 (2016), 257–289. DOI= https://doi.org/10.1515/iral-2015-0022.

[20] Offerijns, J., Verberne, S., & Verhoef, T. 2020. Better Distractions: Transformer-based Distractor Generation and Multiple Choice Question Filtering. *arXiv preprint arXiv:2010.09598*. DOI= https://doi.org/10.48550/arXiv.2010.09598.

[21] R Core Team. 2024. R: *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. DOI= https://www.R-project.org.

[22] Ringlstetter, C., Schulz, K., & Mihov, S. 2006. Orthographic errors in web pages: Toward cleaner web corpora. *Comput. Linguist*. 32, 3 (2006), 295–340. DOI= https://doi.org/10.1162/coli.2006.32.3.295.

[23] Roediger, H. L., & Marsh, E. J. 2005. The positive and negative consequences of multiple-choice testing. *J. Exp. Psychol. Learn. Mem. Cogn.* 31, 5 (2005), 1155–1159. DOI= https://doi.org/10.1037/0278-7393.31.5.1155.

[24] Shin, J., Guo, Q., & Gierl, M. J. 2019. Multiple-choice item distractor development using topic modeling approaches. *Front. Psychol*. 10, Article 825. DOI= https://doi.org/10.3389/fpsyg.2019.00825.

[25] Wickham, H., François, R., Henry, L., & Müller, K. (2024). dplyr: A Grammar of Data Manipulation (*R package version 1.1.4*). DOI= https://CRAN.R-project.org/package=dplyr.

[26] van der Loo, M. P. J. (2024). The stringdist package for approximate string matching (*R package version 0.9.14*). DOI= https://CRAN.R-project.org/package=stringdist.