Issues and concerns in the automatic generation of vocabulary training and testing items

Ralph L. Rose <rose@waseda.jp>
Judy Wang <judy.wang@aoni.waseda.jp>
Naho Orita <orita@waseda.jp>
Ayaka Sugawara <ayakasug@waseda.jp>
Center for English Education in Science and Engineering (CELESE)
Faculty of Science and Engineering, Waseda University

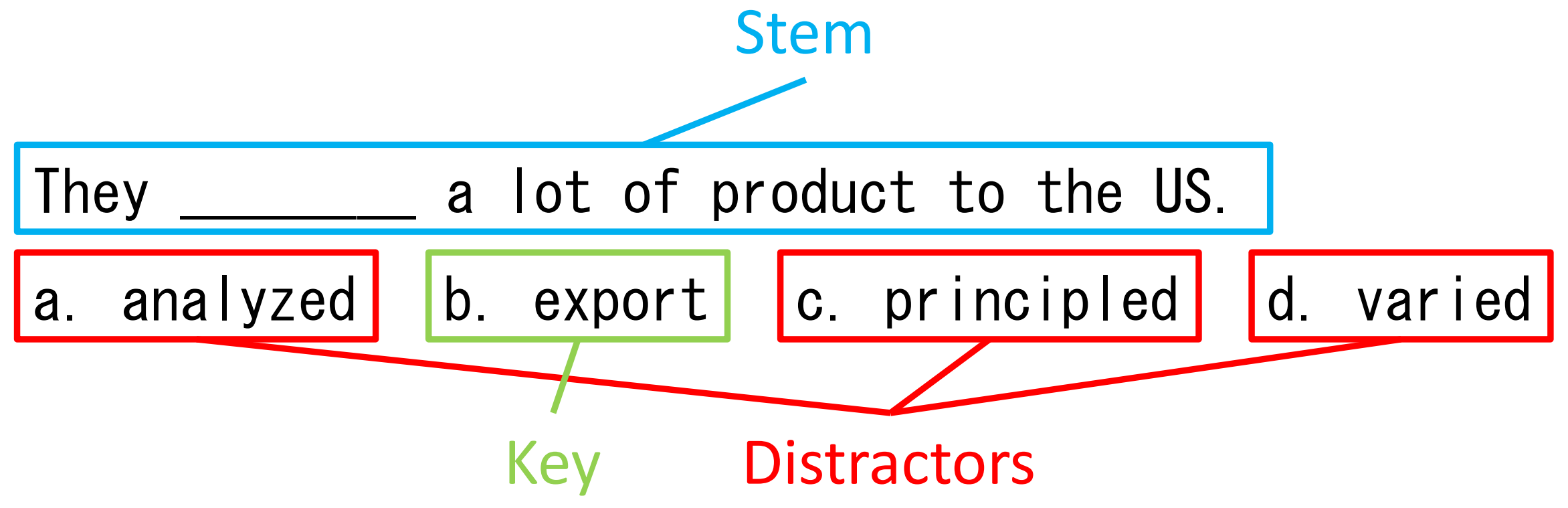
JALTCALL
2022

Abstract

Vocabulary training and testing is an integral part of nearly any language learning program. The VocaTT project aims to make this process easier by building a system for automatically generating items for learners using machine learning algorithms. This progress report focuses on the first stage of this project—to construct a "gold-standard" set of items—and describes issues and concerns in this process. This includes deciding how to generate a large number of items, controlling item difficulty, dealing with sub-standard items, and how learners may interact with the items. The gold-standard set so far contains 2,786 items. These items were used in a pilot experiment with a training and testing application. Participants made modest but definite gains and were motivated to continue their vocabulary study.

Background

Multiple choice cloze (MCC) is widely used in vocabulary testing (Hale et al 1989; inter alia).



Stem

They _______ a lot of product to the US.

a. analyzed b. export c. principled d. varied

Key Distractors

But MCC items present problems in online testing:

- ✗ Labor intensive to produce
- ✗ Not secure (answers easily shared)
- ✗ Cannot be easily re-used

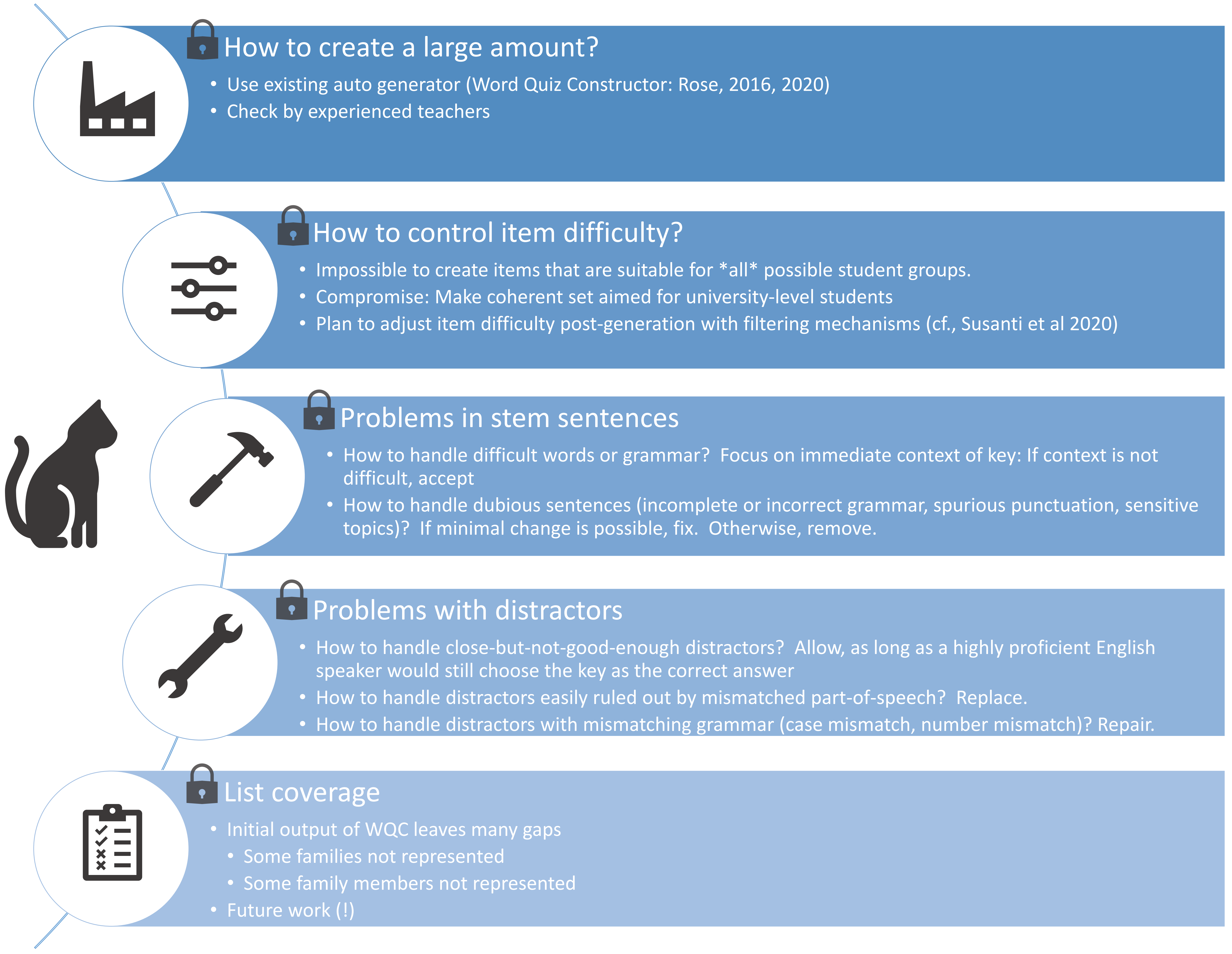
One solution is auto-generation of items en masse. Systems exist for generation of MCC from texts (e.g., Aist 2001; Brown et al 2005; Coniam 1997; Heilman & Eskenazi 2007) or from word lists (Lee et al 2015; Liu et al 2005; Rose 2016, 2020). But few are readily available, easy to use, or adaptable to various needs.

Vocabulary Training & Testing (VocaTT) Project (ongoing) Goals

- ✓ Provide pedagogically sound vocabulary training and testing for learners
- ✓ Provide architecture for large-scale generation of items and extensible for other languages.
- ✓ Generate items for training/testing automatically using a machine learning algorithm trained on "gold standard" items.

Issues and concerns

Constructing a set of gold-standard MCC training and testing items. Objective: generate as large as possible a set of MCC items based on the General Service List (West 1953) and the Coxhead (2000) Academic Word List.



How to create a large amount?

- Use existing auto generator (Word Quiz Constructor: Rose, 2016, 2020)
- Check by experienced teachers

How to control item difficulty?

- Impossible to create items that are suitable for *all* possible student groups.
- Compromise: Make coherent set aimed for university-level students
- Plan to adjust item difficulty post-generation with filtering mechanisms (cf., Susanti et al 2020)

Problems in stem sentences

- How to handle difficult words or grammar? Focus on immediate context of key: If context is not difficult, accept
- How to handle dubious sentences (incomplete or incorrect grammar, spurious punctuation, sensitive topics)? If minimal change is possible, fix. Otherwise, remove.

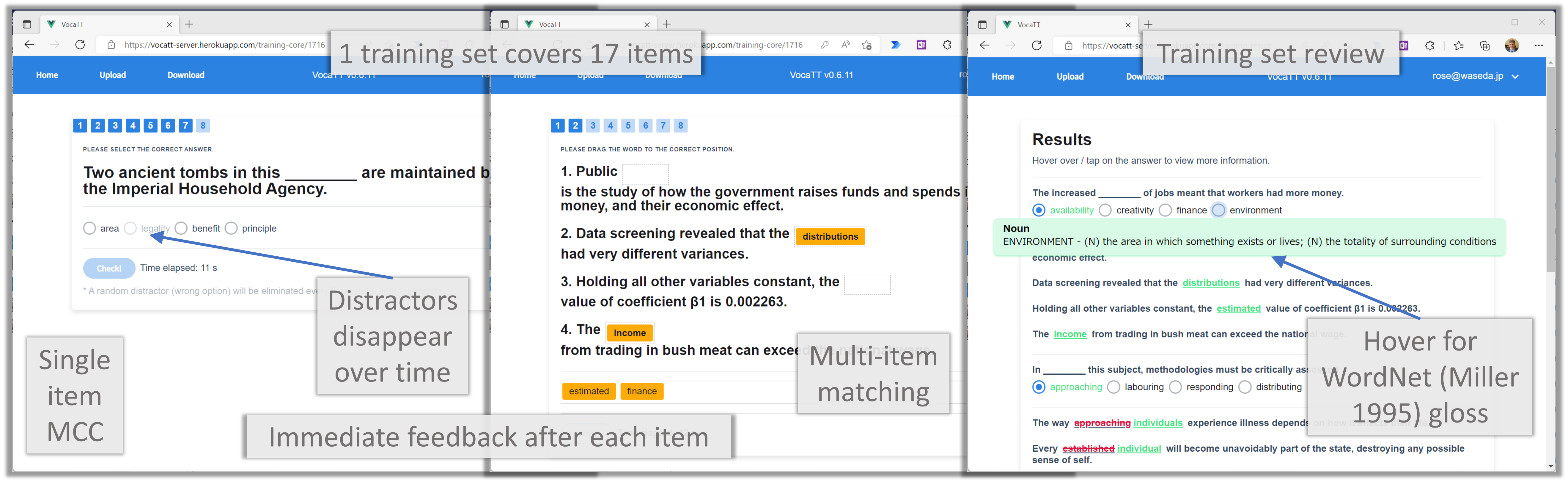
Problems with distractors

- How to handle close-but-not-good-enough distractors? Allow, as long as a highly proficient English speaker would still choose the key as the correct answer
- How to handle distractors easily ruled out by mismatched part-of-speech? Replace.
- How to handle distractors with mismatching grammar (case mismatch, number mismatch)? Repair.

List coverage

- Initial output of WQC leaves many gaps
  - Some families not represented
  - Some family members not represented
- Future work (!)

Providing interface for learners to interact with items. Objective: create simple app with basic training/testing with feedback.



1 training set covers 17 items

PLEASE SELECT THE CORRECT ANSWER.

Two ancient tombs in this _______ are maintained b
the Imperial Household Agency.

○ area ○ legality ○ benefit ○ principle

Check! Time elapsed: 11 s

* A random distractor (wrong option) will be eliminated ev

Single item MCC

Distractors disappear over time

PLEASE DRAG THE WORD TO THE CORRECT POSITION.

1. Public
is the study of how the government raises funds and spends i
money, and their economic effect.

2. Data screening revealed that the distributions
had very different variances.

3. Holding all other variables constant, the
value of coefficient β1 is 0.002263.

4. The income
from trading in bush meat can excee

estimated finance

Multi-item matching

Immediate feedback after each item

Training set review

Results

Hover over / tap on the answer to view more information.

The increased _______ of jobs meant that workers had more money.
● availability ○ creativity ○ finance ○ environment

Noun
ENVIRONMENT - (N) the area in which something exists or lives; (N) the totality of surrounding conditions

economic effect.

Data screening revealed that the distributions had very different variances.

Holding all other variables constant, the estimated value of coefficient β1 is 0.002263.

The income from trading in bush meat can exceed the nation

In _______ this subject, methodologies must be critically ass
● approaching ○ labouring ○ responding ○ distributing

The way ~~approaching~~ individuals experience illness depends

Every ~~established~~ individual will become unavoidably part of the state, destroying any possible sense of self.

Hover for WordNet (Miller 1995) gloss

Pilot test of VocaTT app

The basic suitability of the VocaTT app was pilot-tested with 12 Waseda University students. They completed 60 training and 12 testing sessions over a 2-week period at their convenience.

Results of 30-item test (items not in app)

Pre-test mean (sd)	Post-test mean (sd)	t(11) (p)
19.8 (4.7)	21.6 (5.6)	2.6 (.025)

Results of post-experiment usability survey (4-pt Likert scale; 1=strongly disagree … 4=strongly agree)

Question	Mean	t(11)
I found VocaTT easy to use.	3.1 (0.8)	2.55 (.027)
I found VocaTT fun to use.	3.0 (1.0)	1.82 (.097)
I found VocaTT useful for vocabulary training.	3.2 (1.0)	2.69 (.021)
I would use VocaTT in the future for vocabulary training.	3.1 (0.9)	2.24 (.046)

Acknowledgments

This work was supported by a Waseda University Grant for Special Research Projects (project number: 2021R-026).

References

Aist, G. 2001. Towards automatic glossarization: automatically constructing and administering vocabulary assistance factoids and multiple-choice assessment, International Journal of AI in Ed 12: 212–231.

Brown, J., Frishkoff, G. and Eshkenazi, M. 2005. Automatic question generation for vocabulary assessment. Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, pp. 819–826. Association for Computational Linguistics.

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.

Hale, G. A., Stansfield, C. W., Rock, D. A., Hicks, M. M., Butter, F. A., & Oller, J. W. (1989). The relation of multiple-choice cloze items to the Test of English as a Foreign Language. Language Testing, 6(1), 47–76.

Heilman, M. and Eskenazi, M. 2007. Application of Automatic Thesaurus Extraction for Computer Generation of Vocabulary Questions. Proceedings of Speech and Language Technology in Education (SLaTE), 65–68.

Lee, K., Kweon, S., Kim, H. and 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., and 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.

Miller, G.A. 1995. WordNet: A Lexical Database for English. Communications of the ACM 38(11): 39–41.

Rose, R. 2016. "Automatic Word Quiz Construction Using Regular and Simple English Wikipedia". Proc. of the International Technology, Education and Development Conference (INTED), pp. 8032-8040.

Rose, R. 2020. "Improving the Production Efficiency and Well-formedness of Automatically-Generated Multiple-Choice Cloze Vocabulary Questions". In Proc. of 12th Conf. on Lang. Resources and Eval.(LREC 2020), pp. 7096–7103.

Susanti, Y., Tokunaga, T. & Nishikawa, H. 2020. Integrating automatic question generation with computerised adaptive test. RPTEL 15, 9.

West, M. 1953. A General Service List of English Words. London: Longman, Green and Co.