WQC: A tool for quick automatic word quiz construction

Ralph L. ROSE
<rose@waseda.jp>

Center for English Language Education (CELESE)
Waseda University Faculty of Science and Engineering
Tokyo, Japan

WASEDA

Overview

- Background
  - Testing en masse
  - Automatic test creation
- Word Quiz Constructor
  - Purpose
  - Design
- Qualitative evaluation
- Quantitative evaluation
- Future plans

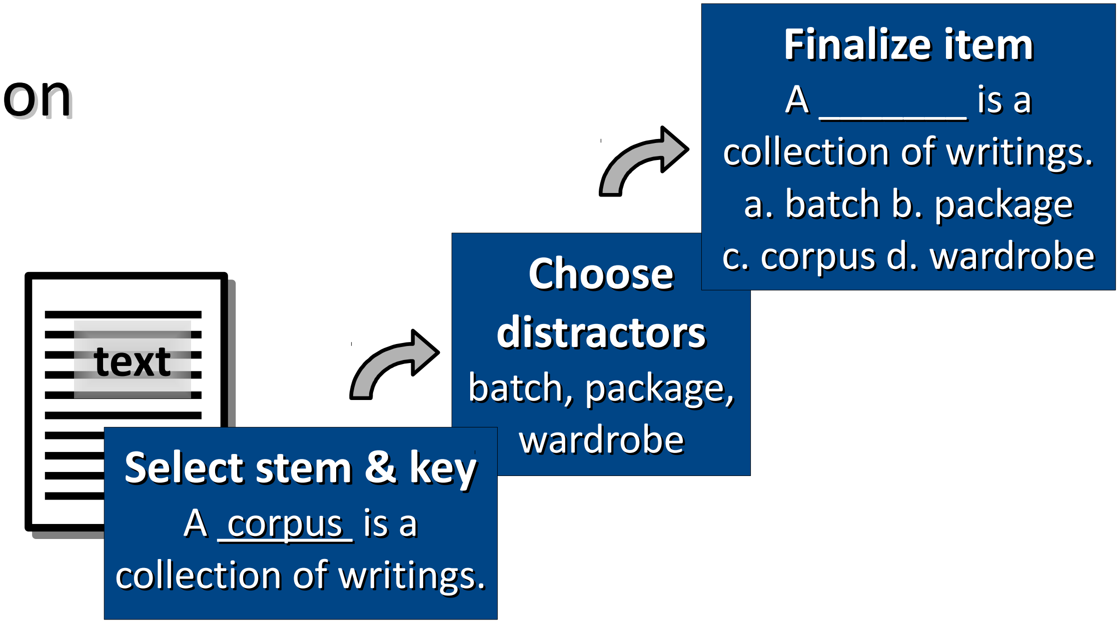
Background: Testing en masse

- Benefits
  - Comparison against large(r) populations; leads to washback effect on instructional planning
  - Minimize test variability; closer adherence to testing objectives (i.e., quality control)
- Challenges
  - Maintaining security over time and location
  - Ensuring competent and consistent administration
  - Ensuring test validity

(Abu-Alhija 2007; Fulcher and Davidson 2007; Weir 2005; inter alia)

Background: Automatic test creation

- Systems
  - Test key concepts (Goto et al 2010; Kunechika et al 2003; Mitkov et al 2006, 2009; Pino et al 2008; Sumita et al 2005)
  - Test vocabulary items in a text (Aist 2001; Brown et al 2005; Coniam 1997; Heilman and Eskenazi 2007)
- Question types
  - Multiple-choice question
  - Multiple-choice cloze
  - Free-response cloze
  - Matching/ordering



Select stem & key
A corpus is a collection of writings.

Choose distractors
batch, package, wardrobe

Finalize item
A ________ is a collection of writings.
a. batch b. package
c. corpus d. wardrobe

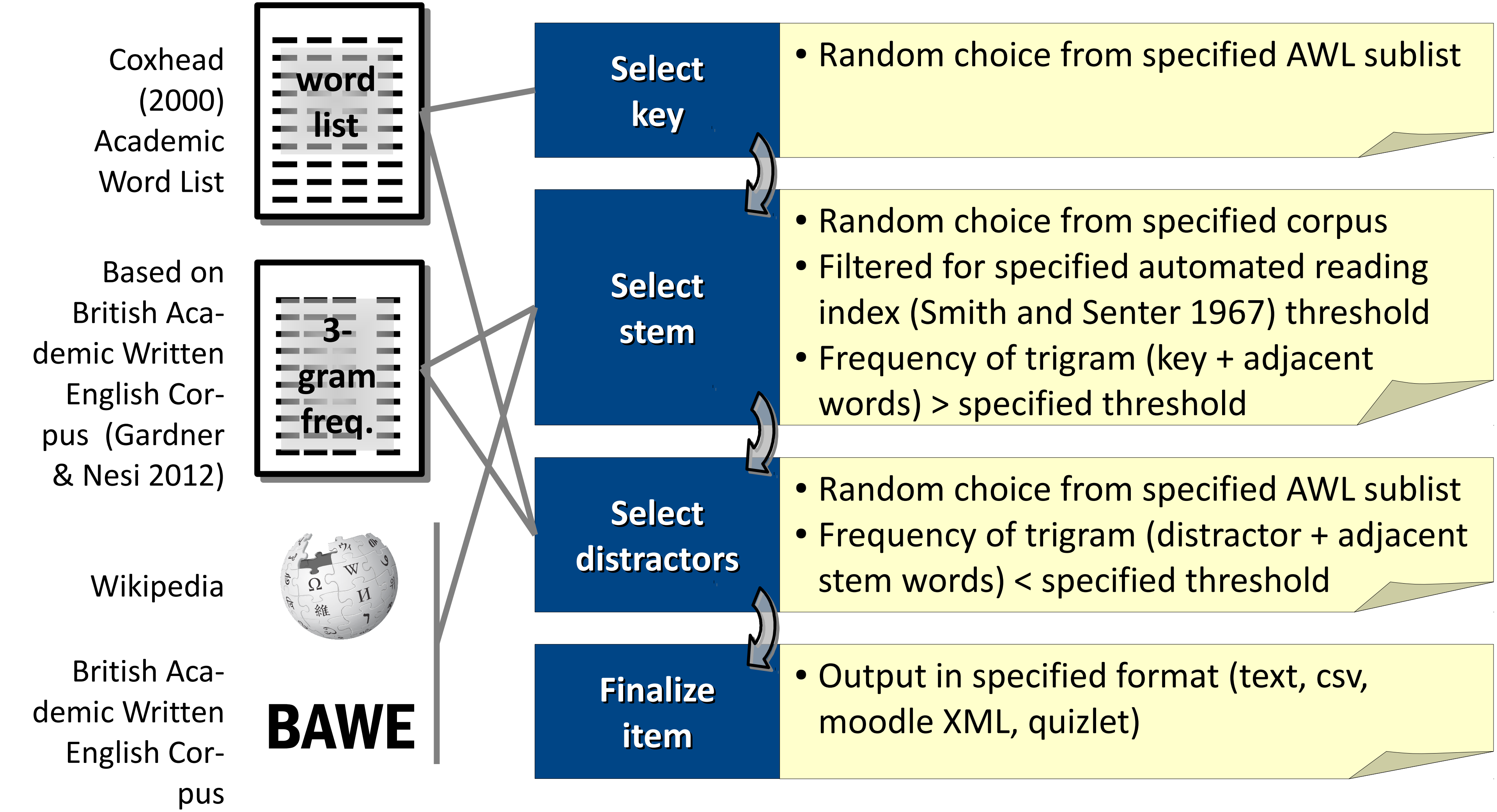
Background: Automatic test creation

- Limitation
  - Input is assumed to be a reading text
  - Many systems are not freely available
- Common vocabulary teaching/learning approach
  - Focus on periodic vocabulary lists
  - Testing targets current list
  - (cf., Brown and Perry 1991; Khoii and Sharififar 2013; Sagarra and Alba 2006)
- Constraints on automated test creation
  - Need a source for stems
  - Key and distractors should be from same list

Waseda CELESE vocabulary testing

- Situation
  - Target vocabulary: Coxhead Academic Word List (2000) sub-lists
  - Study term: One sublist is tested every two weeks
  - Testees: Approx. 2000 students in 50 classes
    - Monday – Friday
    - Morning and afternoon
    - 30 different teachers
- Concerns
  - Teacher preparation time
  - Uniformity (inter-teacher)
  - Consistency (intra-teacher)

Design



Resources

Coxhead (2000) Academic Word List

Based on British Academic Written English Corpus (Gardner & Nesi 2012)

Wikipedia

British Academic Written English Corpus

Procedure

Select key

- Random choice from specified AWL sublist

Select stem

- Random choice from specified corpus
- Filtered for specified automated reading index (Smith and Senter 1967) threshold
- Frequency of trigram (key + adjacent words) > specified threshold

Select distractors

- Random choice from specified AWL sublist
- Frequency of trigram (distractor + adjacent stem words) < specified threshold

Finalize item

- Output in specified format (text, csv, moodle XML, quizlet)

(see Lee et al 2013; Liu et al 2005 for similar approaches)

Design

Sample multiple-choice cloze items

In 2001, 32.4% of the population over the age of fifteen had not completed high school, which is the highest _________ of all three of Saguenay's boroughs. (Wikipedia, ARI=14.6)
a. percentage b. consistency c. derivation d. methodologies

On the local level Benum was _________ in local politics in Verdal municipality from 1959 to 1979. (Wikipedia, ARI=9.2)
a. involved b. constituted c. similar d. uncontextualised

One of the main _________ of decentralisation is the promotion of regional autonomy (Policy guidelines, 2006). (BAWE, ARI=14.5)
a. contexts b. principles c. labors d. illegality

It is measured in the percent rate of real GDP and is considered to be an increase in the _________ of a nation. (BAWE, ARI=7.7)
a. beneficiary b. analyser c. indicators d. income

Evaluation

- Informal usage data
    - Since Spring, 2013
    - Ad-hoc analysis of post-edits
- Controlled comparison of 40 WQC-produced items with 20 manually-produced items
    - Teacher (n=12) judgments of well-formedness and difficulty
    - Student (n=22) response to test items
    - Comments

Qualitative evaluation: Post-editing

- Review of 200 used items (10 test x 20 items/test)
  - 25 items were modified
    - Deleted Wikipedia footnotes in 17 items
    - Added words for clarification/cohesion in 4 items
    - Changed options in 5 items

Critics praised the accurate portrayal of Eddy toward the games, ~~(28) (29)~~ but criticized the _________ of Eddy's role in the film.

As _________ by its name "Les Trois Vallées", the area originally consisted of three valleys: Saint-Bon, Allues, and Belleville.

The _________ was previously known as the "Ministry of Justice and Ecclesiastical Affairs".

a. categorisation b. ~~residence~~ c. resources d. institution
b. participation

Qualitative evaluation: Teacher comments

- Problematic items
  - “In many of the questions it seems that students would be able to guess the correct answer simply by eliminating the distractors that do not fit grammatically in the sentence (i.e. focusing on form, rather than meaning). Thus they may not actually be testing the student's knowledge of the word.”

For the first time in the Soviet Union developed a _________ of chemical defense in the fight against this disease.

a. ~~responses~~ b. ~~contracts~~ c. ~~periods~~ d. method

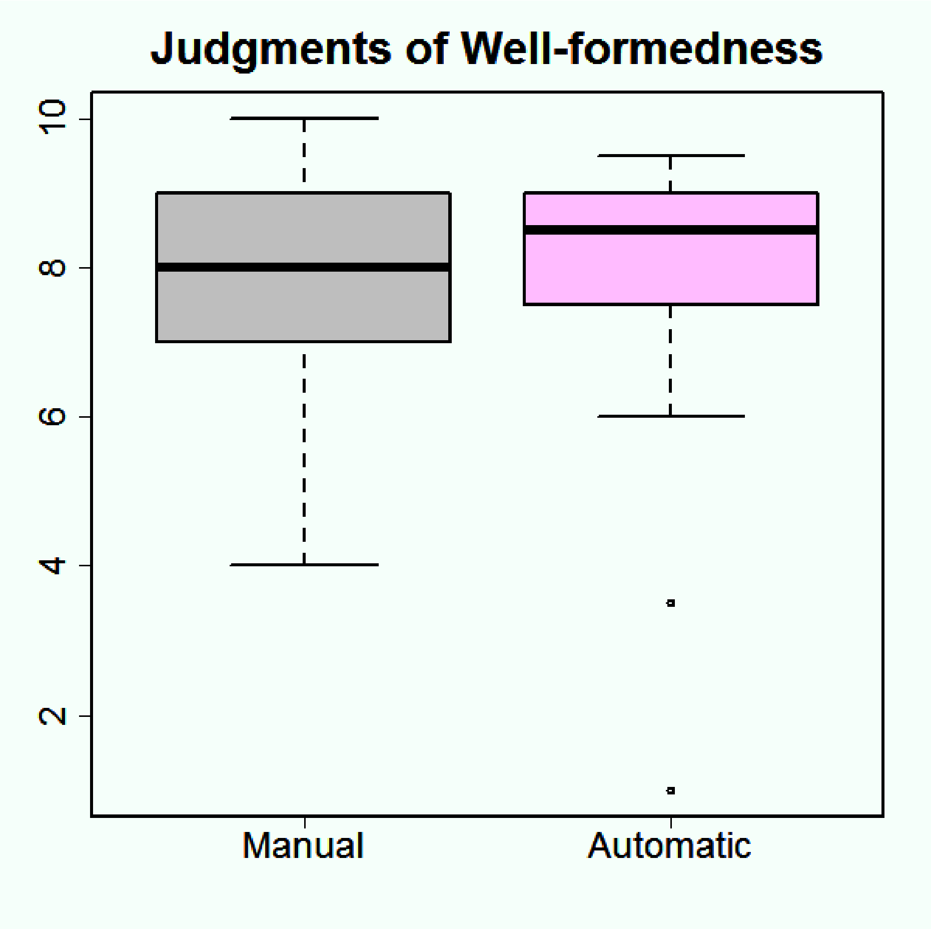
~~(2) (3) (4)~~ He was a member of the Constituent Assembly of India in 1948 from Bihar.

a. occurrences b. formulated c. constituent d. evidential

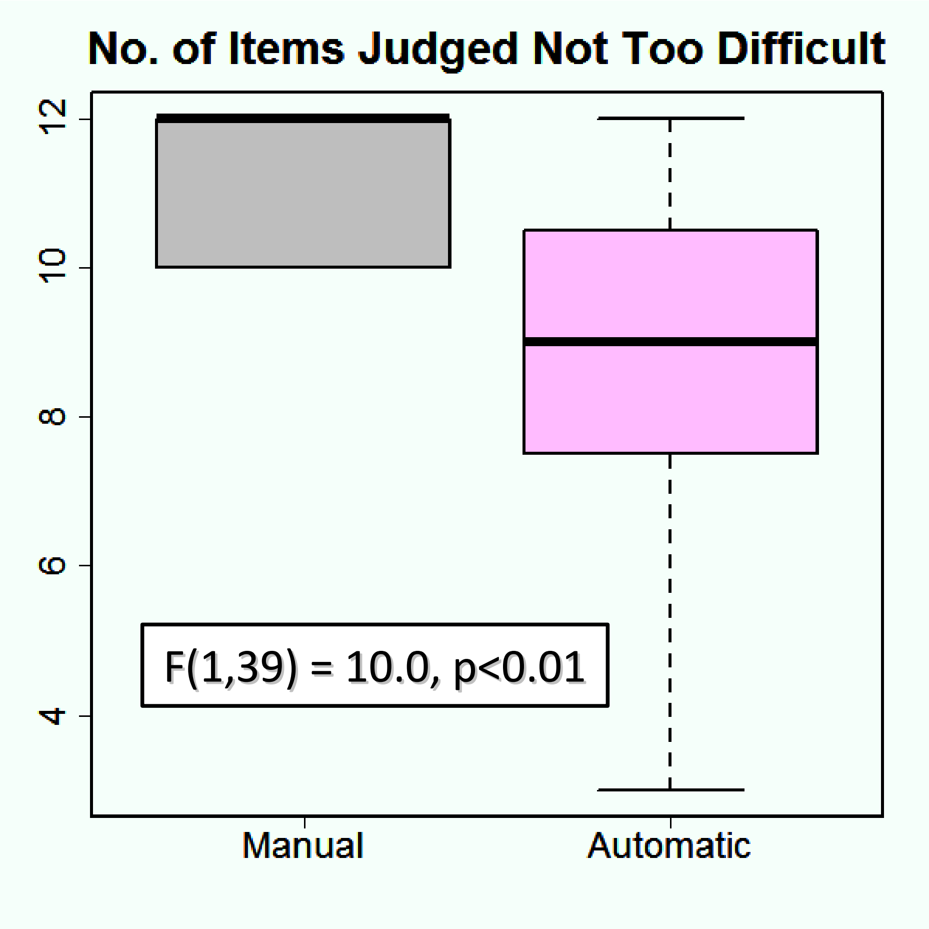
Qualitative evaluation: Teacher comments

- Items coincide with intuitions about frequency
  - “I did have one student challenge an item, saying he didn't think the correct answer choice collocated with the word next to it. I knew it did, but I checked it with COCA and gave him a print out of the top 20 most frequent collocations (it appeared in that list).”

Quantitative evaluation: Teacher judgment



Judgments of Well-formedness

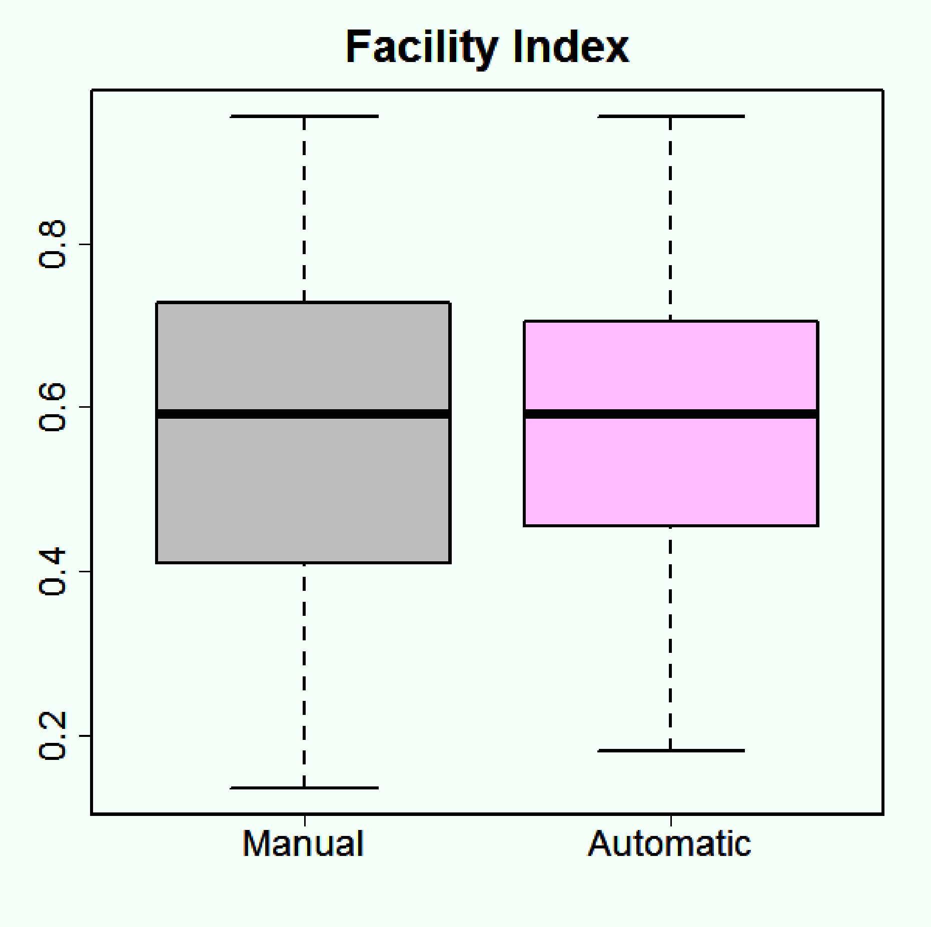


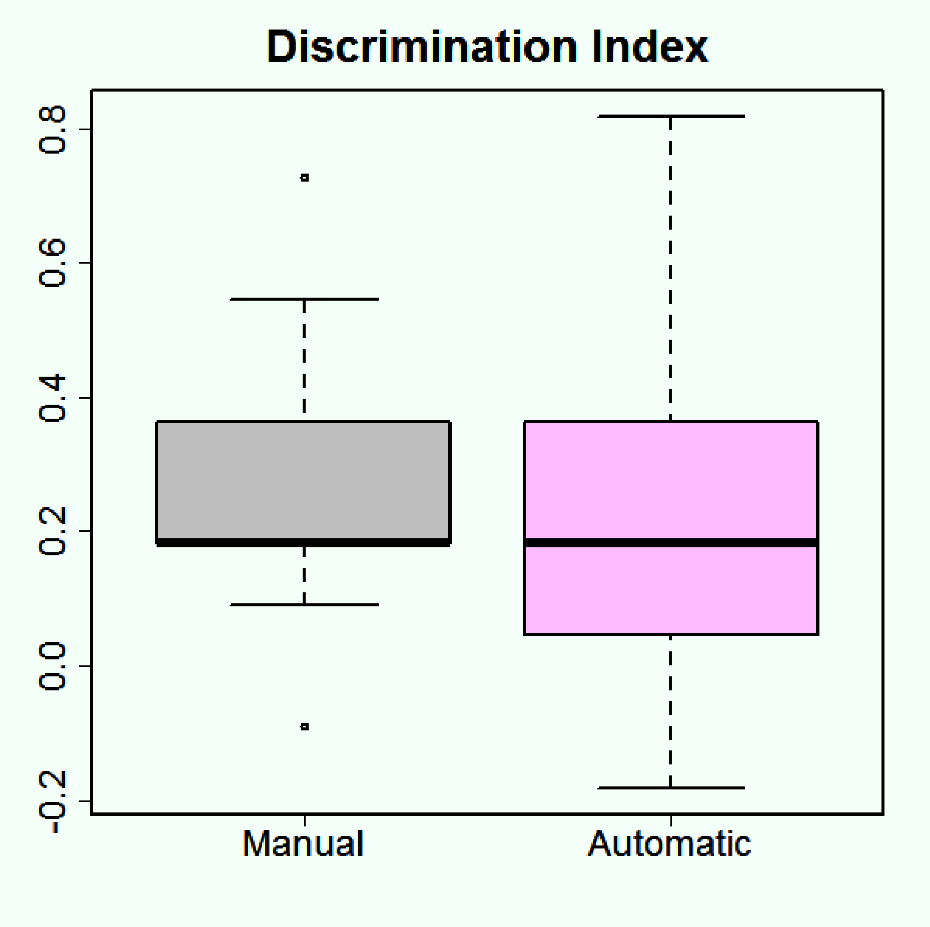
No. of Items Judged Not Too Difficult

F(1,39) = 10.0, p<0.01

Students also commented about difficulty of items.

Quantitative evaluation: Student response

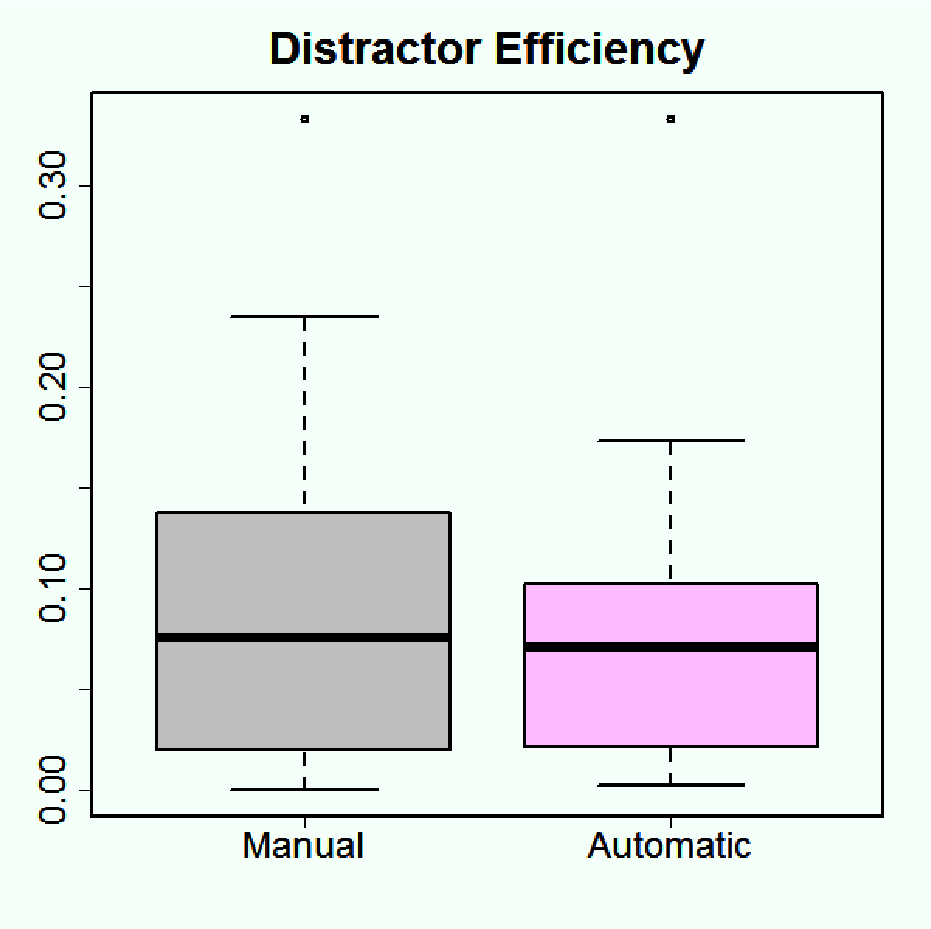


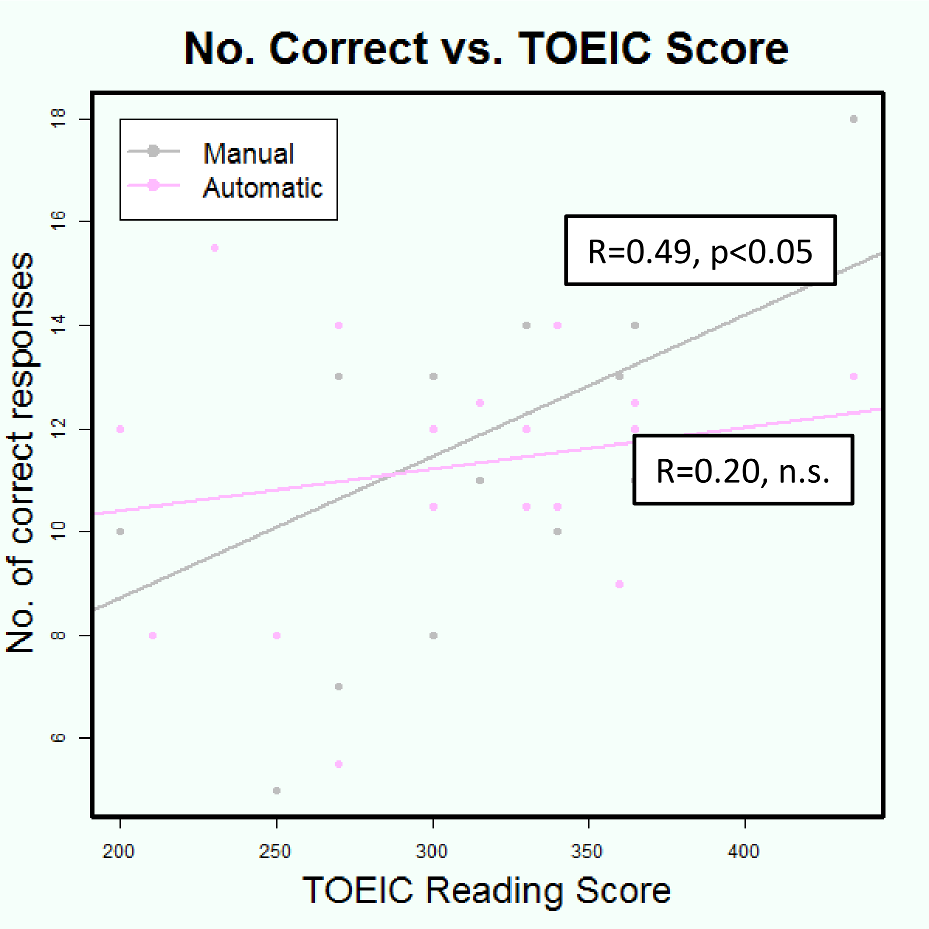


Student impression of difficulty is not unique to automatic items.

Target range is usually 0.3 to 0.7.

Quantitative evaluation: Student response


Distractor Efficiency


No. Correct vs. TOEIC Score

Variance in the distribution of distractor selection. Target efficiency is <0.08.

Summary of findings

- WQC can produce test items that are comparable to those produced manually.
- Problematic items result primarily from breaking textual cohesive links or impossible distractors.
- Divergence between teachers and students in terms of apparent difficulty and actual difficulty.
- Teachers can be confident that the contexts of correct answers are high frequency sequences.

Future work

- Add other question types (e.g., matching, word-ordering).
- Construct a graphical user interface.
- Expand capability for other vocabulary lists.
- Prepare application for free distribution.

References

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

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.

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. 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. Proceedings of Speech and Language Technology in Education (SLaTE), 65-68.

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

Kunichika, H., Katayama, T., Hirashima, T. and Takeuchi, A. 2003. Automated question generation methods for intelligent English learning systems and its evaluation. Proceedings of ICCE2004.

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.

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. Proceedings of the EACL 2009 Workshop on GEMS: GEometical Models of Natural Language Semantics, 49-56.

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, 22-32.

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. Proceedings of the 2nd Workshop on Building Educational Applications Using NLP, 61-68.

Weir, C.J. 2005. Language Testing and Validation: An Evidence-based Approach. Palgrave-Macmillan.