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
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 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

Choose distractors
batch, package, wardrobe

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

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

• 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**

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

**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

Students also commented about difficulty of items.

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

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

Distractor Efficiency

No. Correct vs. TOEIC Score

R=0.49, p<0.05

R=0.20, n.s.
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.