
Ralph L. ROSE
<rose@waseda.jp>

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

International Technology, Education, and Development Conference (INTED)
7-9 March 2016
Valencia, Spain
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
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
Word Quiz Creator (WQC) design

Resources

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

Procedure

1. **Select key**
   - Random choice from specified AWL sublist

2. **Select stem**
   - Random choice from specified corpus
   - Filtered for specified automated reading index (ARI: Smith and Senter 1967)
   - Frequency of trigram (key + adjacent words) > specified threshold

3. **Select distractors**
   - Random choice from specified AWL sublist
   - Frequency of trigram (distractor + adjacent stem words) < specified threshold

4. **Finalize item**
   - Output in specified format (text, csv, moodle XML, quizlet)

(see Lee et al 2013; Liu et al 2005 for similar approaches)
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
Previous work with WQC

- WQC can produce test items comparable to manual items: facility, discrimination, distractor efficiency, and face validity with teachers (Rose 2014a, 2014b)
- However, stems from Wikipedia were regarded by teachers and students as rather difficult or long.
  - Chemical symbols and abbreviations as short words
    
    E-MR1s are ________ in matte silver or matte olive. (ARI=5.57)
    a. available  b. resourceful  c. complex  d. normal

  - High ARI threshold allows difficult technical words
    
    Also, messages in the Actor model are simply sent (like packets in IP); there is no ________ for a synchronous handshake with the recipient. (ARI=13.92)
    a. sectors  b. derivations  c. requirement  d. significance
Simple English Wikipedia

- Wikipedia has many language variants
  - Japanese, Russian, Hindi, Swahili, ...
  - English and Simple English

- Editorial advice for Simple English page writers (Wikipedia contributors 2016):
  - "...should use only the 1,000 most common and basic words in English"
  - "...simple grammar and shorter sentences."

- Hypothesis: Simple English pages would provide a more reliable source of stems than regular English pages.
Experiment 1: Quantitative comparison

• Multiple-choice cloze items for comparison (from AWL sublists 1 & 2)
  - WQC (using ARI threshold ≤ 16)
    • 400 items using regular English Wikipedia
    • 400 items using Simple English Wikipedia
  - Manually-produced
    • 30 items produced by experienced ES/FL instructor
      (previously used in classroom testing in Japan university-level EFL instruction)

• Evaluated:
  - Time to produce
  - Readability (via ARI)
  - Length
Experiment 1: Quantitative comparison

- SEW items produced faster than EW items
  - EW: 67.4 sec/item    SEW: 30.7 sec/item
- SEW items more readable than EW items; comparable to Manual items.

**Word length**

- $F(2,827)=12.1$, $p<0.001$

**Char length**

- $F(2,827)=15.3$, $p<0.001$

**ARI**

- $F(2,827)=16.2$, $p<0.001$

Boxplots produced in R: Dark black line indicates median; shaded regions represent 2nd and 3rd quartiles.
Experiment 2: Native validation

- Amazon Mechanical Turk
  - Workers do on-line Human Intelligence Tasks (HITs) for remuneration.
  - Used by more and more linguistics researchers (Schnoebelen and Kuperman 2010)
- Multiple-choice test with 120 items
  - First 40 regular English Wikipedia items from Expt. 1
  - First 40 Simple English Wikipedia items from Expt. 1
  - 30 manual items from Expt. 1
  - 10 pre-validated “check” items to assure good work from workers (excluded from analysis).
- HIT completed by 51 workers; 1 worker’s results excluded because check items were incorrect.
Experiment 2: Native validation

- Split items into low and high groups by ARI

<table>
<thead>
<tr>
<th></th>
<th>EW</th>
<th>SEW</th>
<th>Man</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>12.3</td>
<td>10.7</td>
<td>13.3</td>
</tr>
<tr>
<td>Low</td>
<td>7.8</td>
<td>5.9</td>
<td>6.5</td>
</tr>
</tbody>
</table>

- Facility index (proportion of correct responses) is consistently best for regular English Wikipedia items; slightly diminished for high level Simple English items.

Facility index

F(2, 104) = 3.26, p < 0.05
Discussion and future plans

• Discussion
  - Is SEW better than EW for WQC item generation?
    • Yes, it’s faster, and item stems are shorter and more readable.
    • No, higher level items are diminished in facility.
  - Use SEW with low ARI threshold (e.g., ≤10); but production time will increase

• Future plans
  - Evaluate SEW items with nonnative testees
  - 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


