Automatic Word Quiz Construction Using Regular and Simple English Wikipedia

Ralph L. ROSE
<rose@waseda.jp>

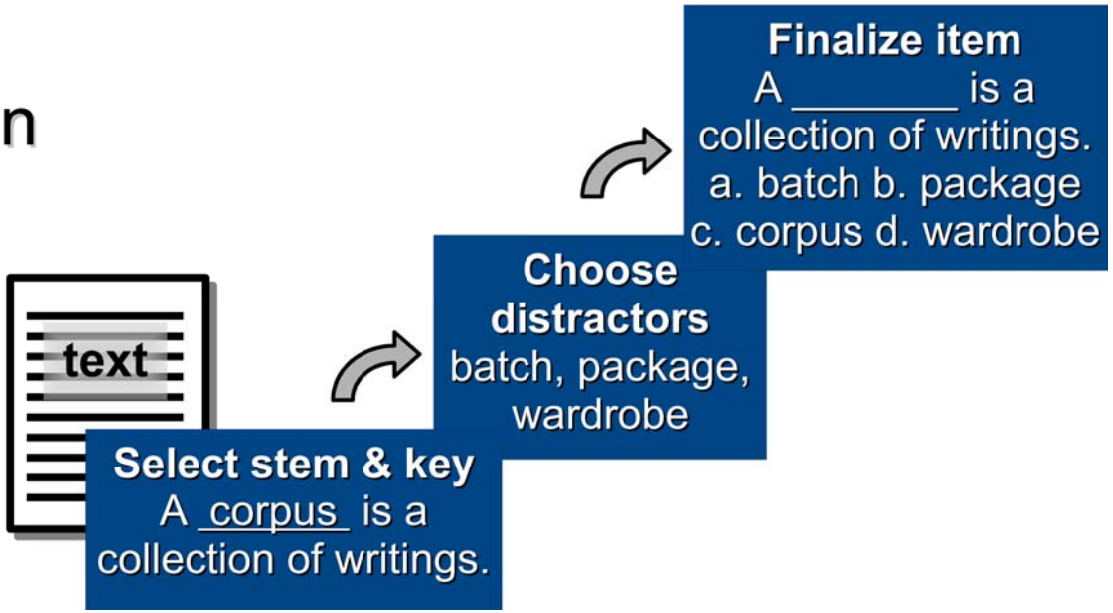
Center for English Language Education (CELESE)
Waseda University Faculty of Science and Engineering
Tokyo, Japan

International Technology,
Education, and Development
Conference (INTED)

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
  - (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

Word Quiz Creator (WQC) design

Resources

Coxhead (2000) Academic Word List

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 (ARI: Smith and Senter 1967)
- 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)

Word Quiz Creator (WQC) 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

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

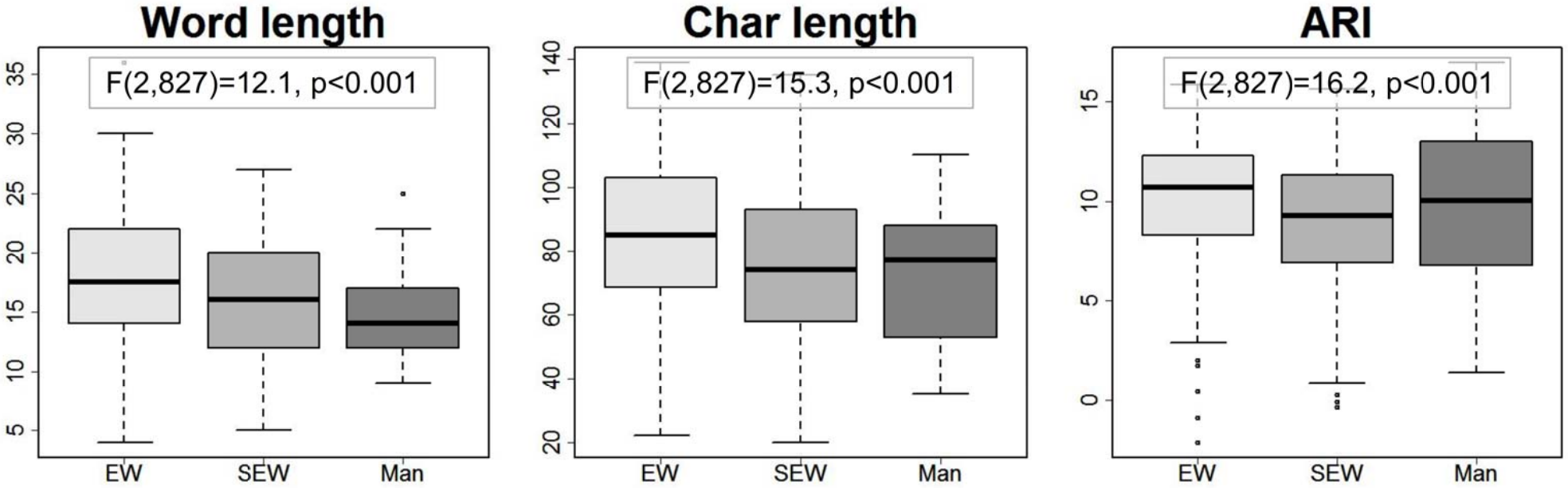
- Wikipedia has many language variants
  - Japanese, Russian, Hindi, Swahili, ...
  - English and Simple English
- Editorial advice for Simple English page writers (Wikipedia contributors 2016):
  - "...should use only the 1,000 most common and basic words in English"
  - "...simple grammar and shorter 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.

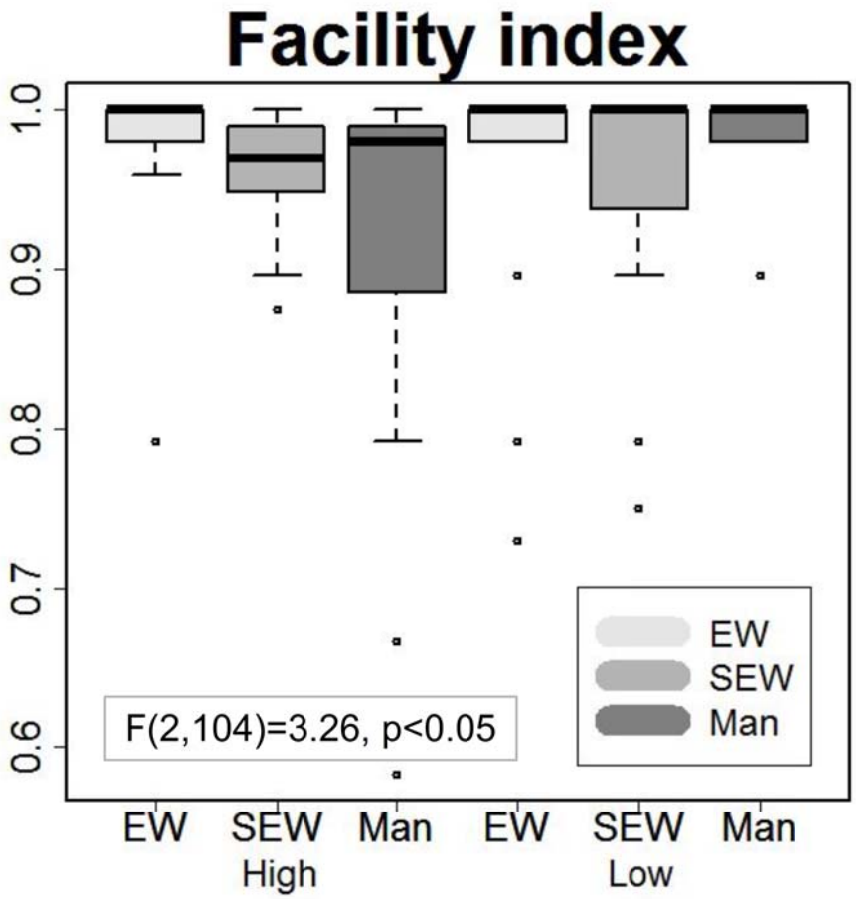


Boxplots produced in R: Dark black line indicates median; shaded regions represent 2nd and 3rd quartiles.

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

	EW	SEW	Man
High	12.3	10.7	13.3
Low	7.8	5.9	6.5

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

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

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.

Wikipedia contributors, "Wikipedia:Simple English Wikipedia," Wikipedia, The Free Encyclopedia, https://simple.wikipedia.org/w/index.php?title=Wikipedia:Simple_English_Wikipedia&oldid=5276509