Automated vocabulary quiz creation using online and offline corpora

TaLC 11
Teaching and
Language Corpora
UCREL at Lancaster U.

Ralph L. ROSE <rose@waseda.jp>, Center for English Language Education (CELESE), Waseda University Faculty of Science and Engineering

1. Abstract

Word Quiz Constructor (WQC) is a Java application designed to create a large number of quizzes from word lists by drawing test materials from online or offline corpora. Experiments with teachers shows that WQC can reliably generate well-formed items on a par with manually produced items. Furthermore, tests with students show that items produced with WQC have facility, discrimination, and distractor effectiveness that is comparable to that of manual items. Finally, results show that online corpora (Wikipedia) are more suitable to producing items that are at a higher reading level.

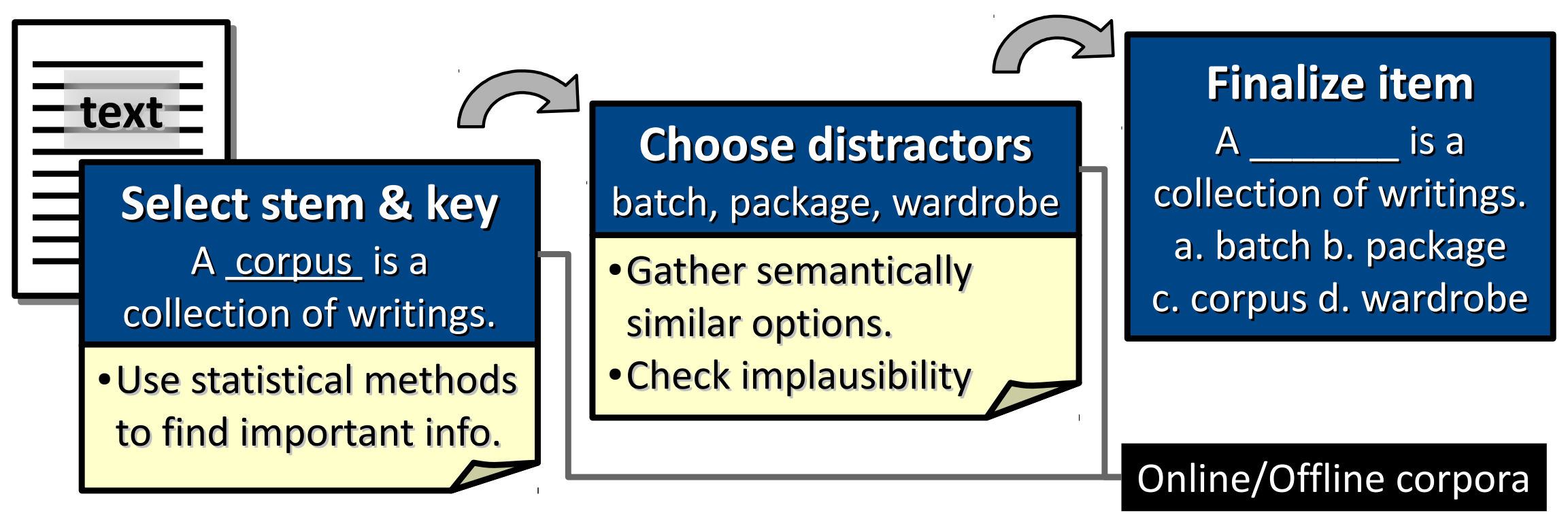
2. Background

Many types of questions may be used to test and evaluate students' knowledge.

Multiple choice question	Multiple choice cloze	Free response cloze	Matching, ordering	Free response (short/long)

Multiple-choice cloze is the most studied and is often used in the automatic construction of test questions (e.g., Goto et al 2010; Kunechika et al 2003; Mitkov et al 2006, 2009; Pino et al 2008; Sumita et al 2005). Most systems rely on various online and offline corpora (BNC, Wikipedia, Google, Wordnet, etc.)

Automatic Construction of Multiple-choice Cloze Questions



text

Select stem & key
A _corpus_ is a collection of writings.
- Use statistical methods to find important info.

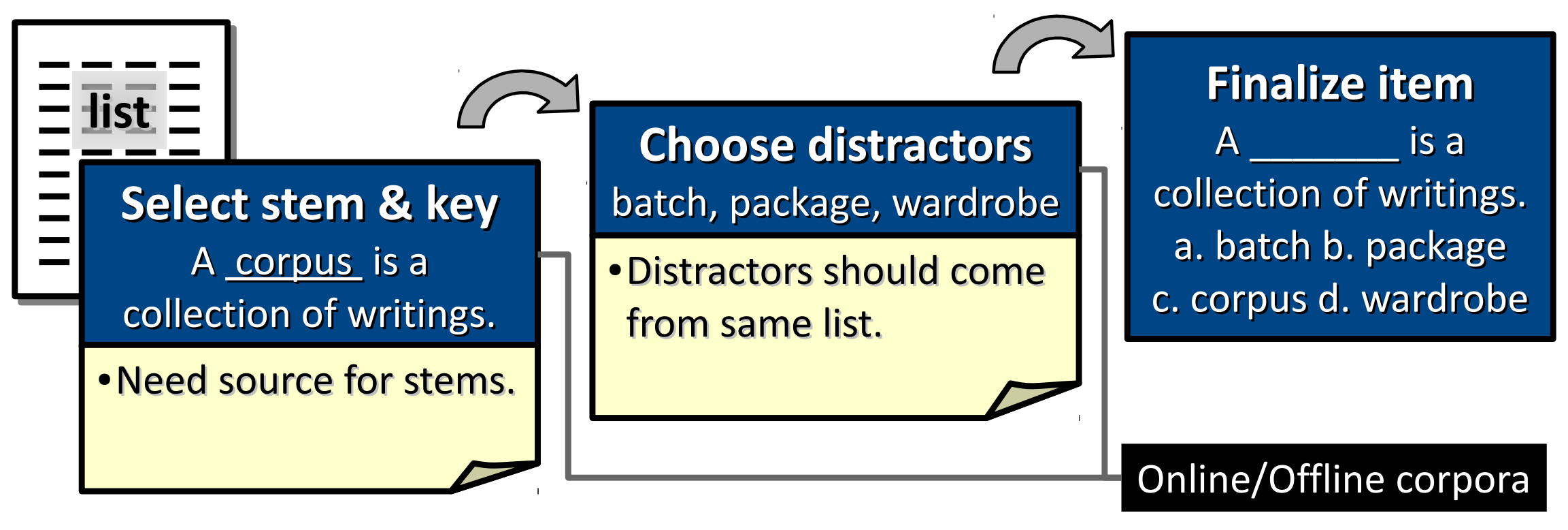
Choose distractors
batch, package, wardrobe
- Gather semantically similar options.
- Check implausibility

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

Online/Offline corpora

The procedure can be used to create cloze questions testing comprehension of key ideas (see citations above) or knowledge of vocabulary items (e.g., Aist 2001; Brown et al 2005; Coniam 1997; Heilman and Eskenazi 2007) appearing in a text. However one time-honored method of vocabulary instruction, training, and testing involves the use of fixed, periodic lists of vocabulary items. Many applications are not readily compatible with this approach (though Lee et al 2013 and Liu et al 2005 are close).

Automatic Construction of Multiple-choice Cloze Questions



list

Select stem & key
A _corpus_ is a collection of writings.
- Need source for stems.

Choose distractors
batch, package, wardrobe
- Distractors should come from same list.

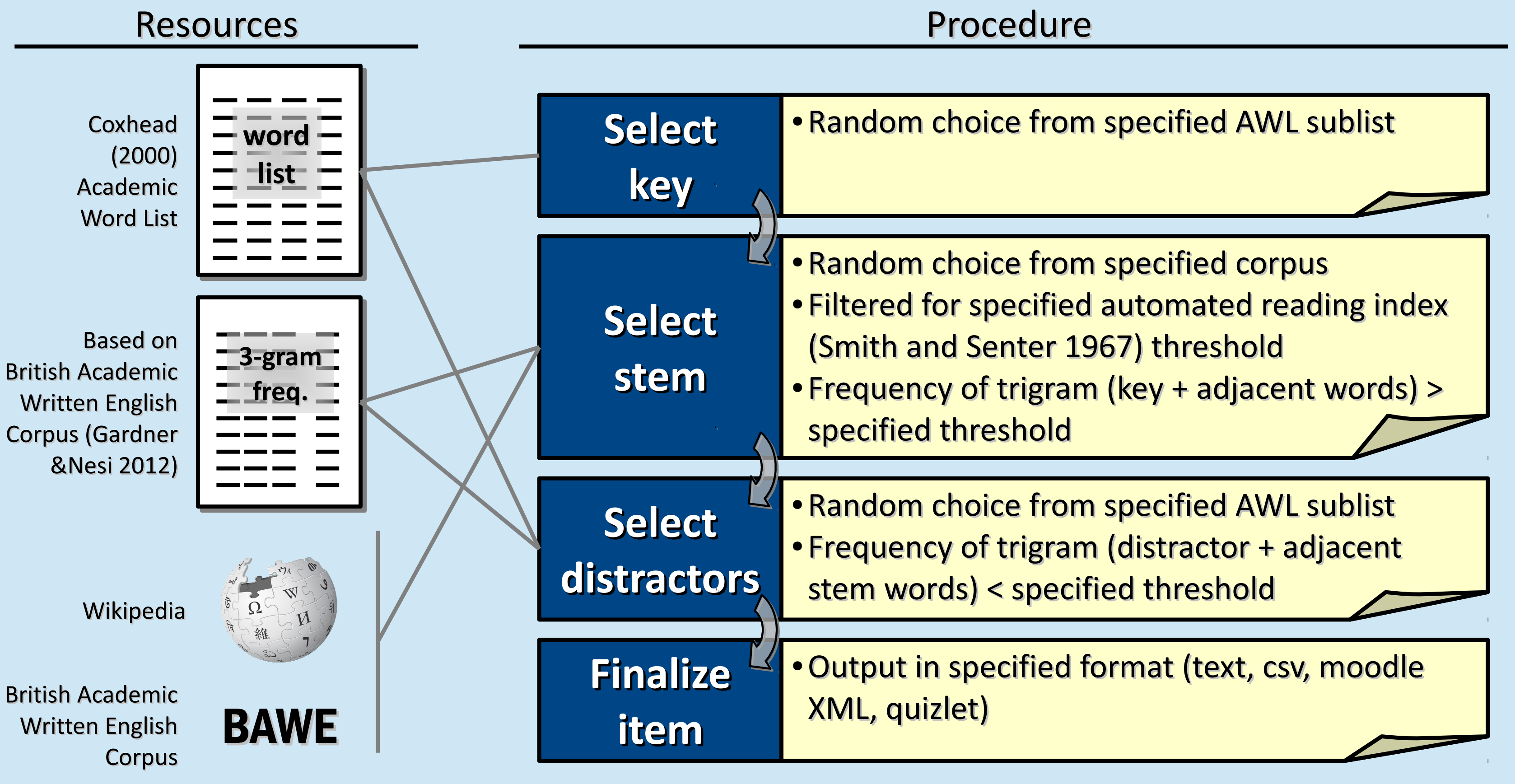
Finalize item
A ________ is a collection of writings.
a. batch b. package
c. corpus d. wardrobe

Online/Offline corpora

Aims of this research project:

- Develop a tool to produce vocabulary quizzes from lists.
- Evaluate quiz items produced using the system with feedback from teachers and students.
- Compare the effectiveness of online and offline corpus resources in the production of quiz items.

3. Word Quiz Creator



Resources

Coxhead (2000) Academic Word List — word list

Based on British Academic Written English Corpus (Gardner &Nesi 2012) — 3-gram freq.

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)

Sample Output

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. methodologies
c. derivation d. consistency

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

4. Experimental Evaluation

Experiment 1: Well-formedness and difficulty

Experienced EFL teachers (N=12, avg teaching exp=21 yrs) judged well-formedness and then difficulty (relative to key) of well-formed items in forced-choice paradigm.

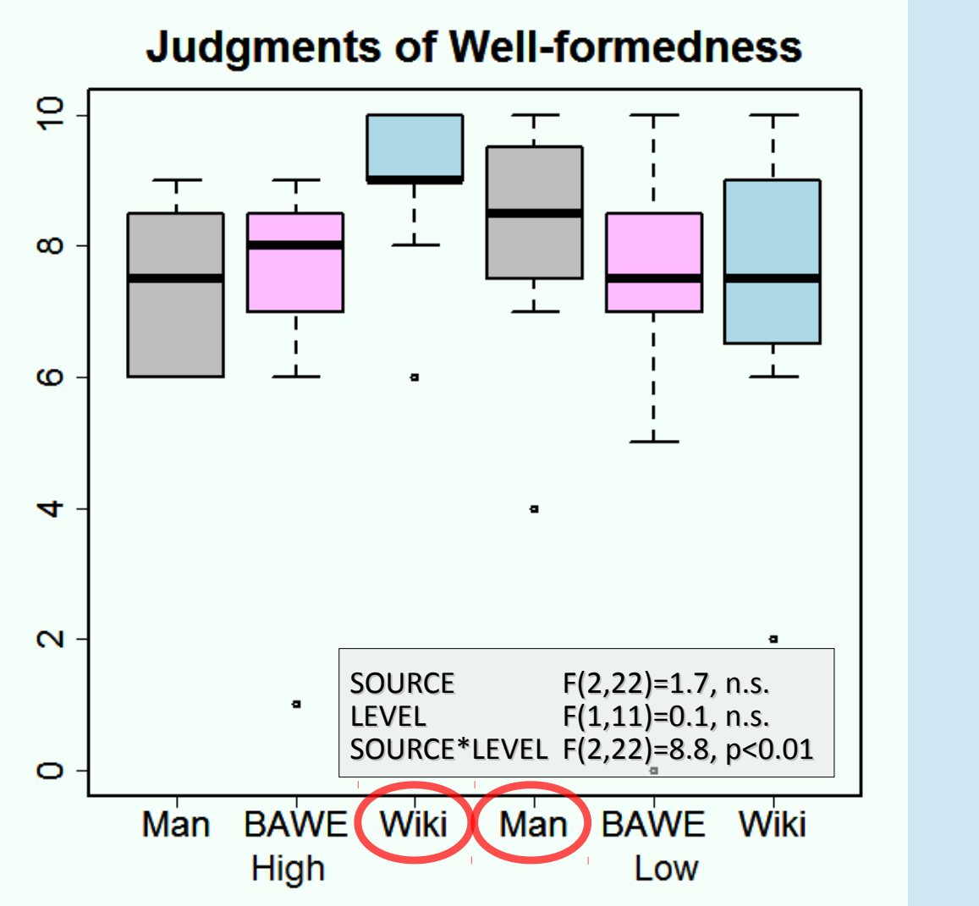
Multiple-choice cloze stimuli

	LEVEL	Low (ARI<12)	High (12<ARI<16)
SOURCE	Manual*	10	10
	BAWE	10	10
	Wikipedia	10	10

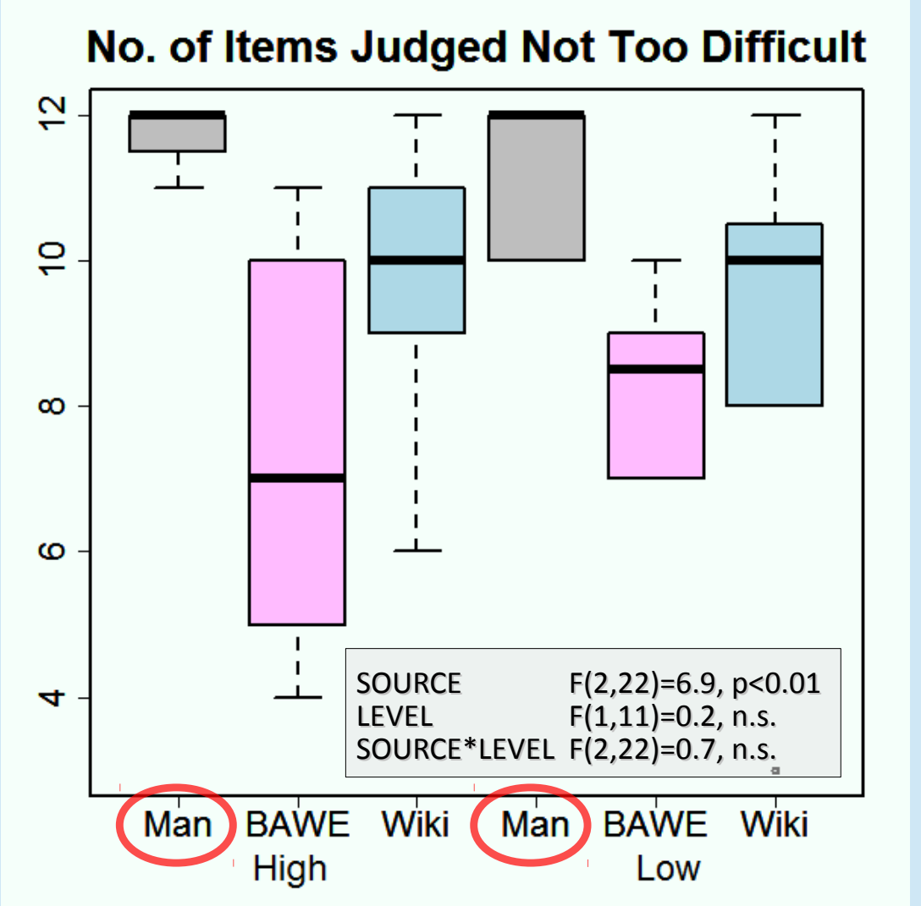
*Manual items prepared by experienced teacher and used in actual teaching.

Experiment 2: Facility, discrimination, efficiency

Undergraduate and graduate university students (N=22) responded to the stimuli items in a simulated vocabulary test in exchange for 1,000 yen (£6) each.



For high ARI, more Wikipedia items are judged well-formed, while for low ARI, more manual items are judged well-formed.



Manual items were consistently judged not too difficult; Wikipedia items were next; followed by BAWE items:

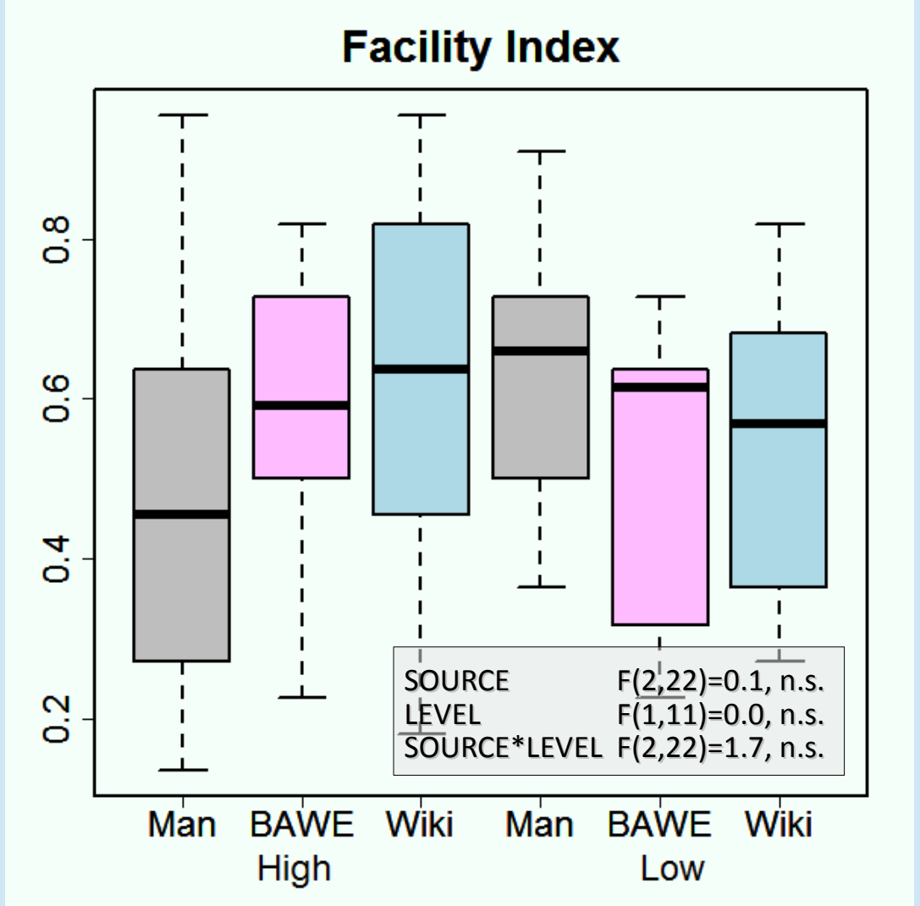
Manual > Wikipedia > BAWE

Problematic items identified by teachers

There are ________ and differences in the history and current meanings of Hosay across the Caribbean. a. ~~illegality~~ b. ~~environmentalist~~ c. ~~indicator~~ d. similarities

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

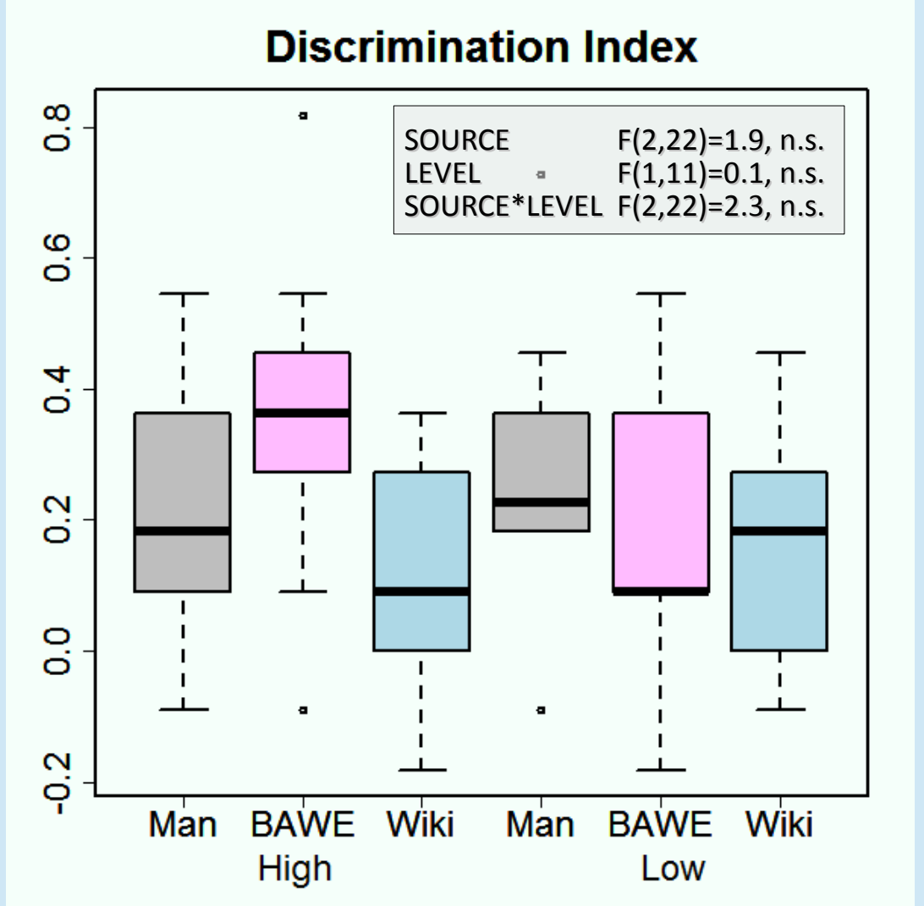


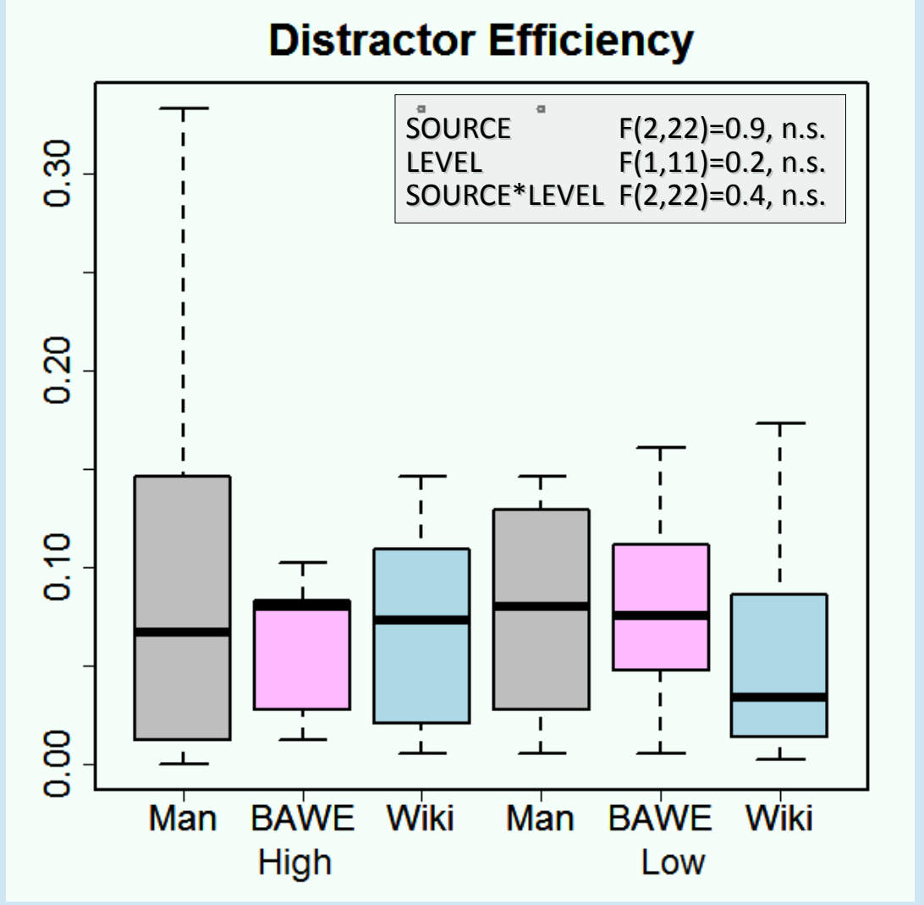
There were no significant differences between groups: Students performed similarly with all items.

Mean facility index of items was in an acceptable range.

Mean discrimination index is positive, but not very high.

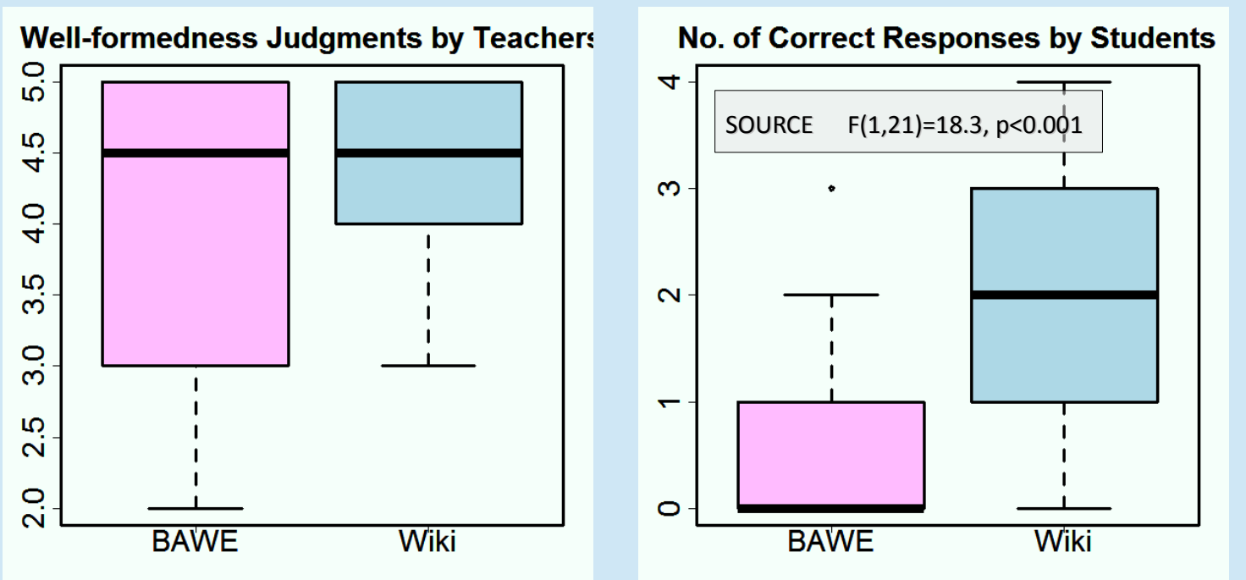
Distractors are minimally efficient (<0.08)





Free-response cloze items

(a) On the supply side, it is ______ that lenders maximise expected profits in a competitive market.
(b) The levels in AB could be ______ to be too low to induce pharyngeal cell fates.
Hint: This word begins 'a' and can be defined as 'take to be the case or to be true'
Answer: _assumed_

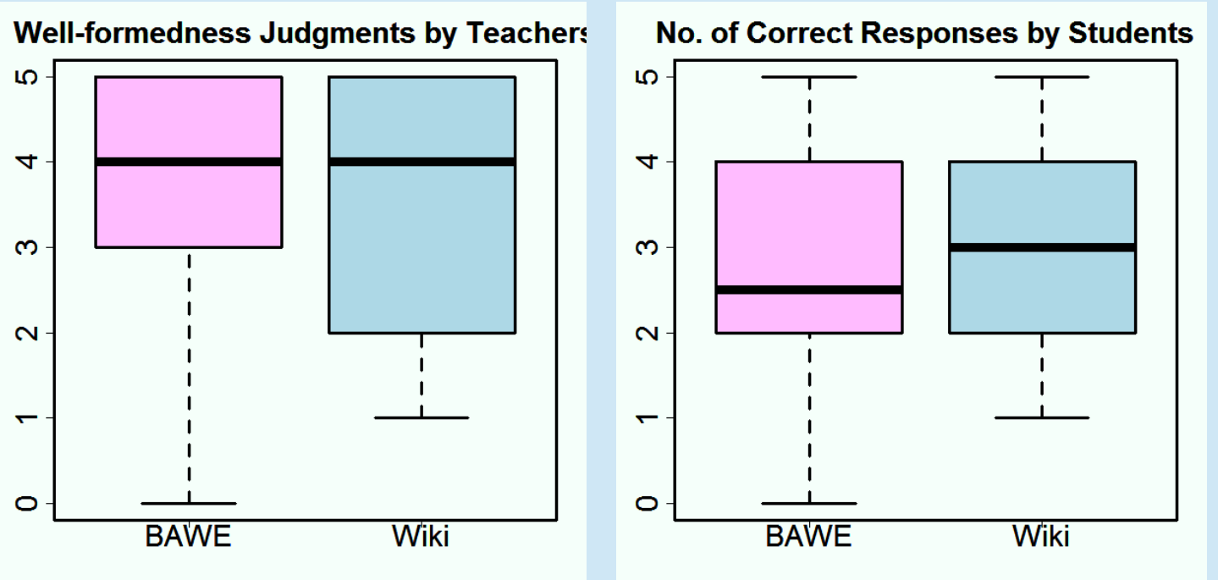


Well-formed but low facility, esp. BAWE items.

Multiple-choice synonym items

The photo and data are {processed} and a physical card is printed on hi-quality photo paper and sent from the USA to any destination.
Which of the following words is closest in meaning to the root word 'process'?
a. authoritative b. contracted
c. treat d. constituted

(glosses from Wordnet: Miller 1995)



Well-formed and moderate facility.

5. Summary

How do the BAWE and Wikipedia items compare to manual items?

	BAWE	Wikipedia
Well-formedness	same	high better
Difficulty	much worse	a little worse
Facility	same	same
Discrimination	same	same
Distractor efficiency	same	same
Prep time (ARI<16)	25 sec/item	117 sec/item

6. Discussion

One possible reason for the well-formedness drop in low Wikipedia items may be that Wikipedia's writing style is normally quite high. Items with low ARI might not be normal writing: heavy in abbreviations, footnotes, or academic shorthand (e.g., math equations). It may be useful to have a lower ARI threshold in addition to the upper threshold to control this. Wikipedia items also take significantly longer to produce, but that is probably caused by networking delays and limits on the number of API requests by the Wikipedia server.

The limitations with BAWE likely result from the smaller size of the corpus: Many items may fail to be finalized, thus costing time. Those that are finalized are more difficult, perhaps because of a greater concentration of difficult vocabulary. Nonetheless, on the whole, Word Quiz Creator is capable of producing vocabulary test items on a par with those produced manually. An online corpus (Wikipedia) and an offline corpus (BAWE) perform somewhat variably in this process, but may potentially complement each other to produce useful items.

7. Future work

Intended improvements to Word Quiz Creator include:

- Use Google n-grams rather than BAWE n-grams. This should increase item acceptability rate, speeding up production time.
- Use a local server installation of Wikipedia rather than hit the Wikipedia site directly, also speeding up production time.
- 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.

Furthermore, since the current study's scale is relatively small, future work will include more extensive testing of the Word Quiz Creator's output in order to validate the usefulness of the items for testing vocabulary knowledge.

Acknowledgments

This research is partially supported by a Waseda University Grant for Special Research Projects (Project number: 2014K-6142). I am also grateful to Laurence Anthony for advice and assistance at the beginning stage of this work and also to several teachers who participated voluntarily in Experiment 1.

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.

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.

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.

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.