EXERCISE MAKER: АВТОМАТИЧЕСКОЕ СОЗДАНИЕ ЯЗЫКОВЫХ УПРАЖНЕНИЙ

Малафеев А. Ю. (aumalafeev@hse.ru)

Национальный исследовательский университет
«Высшая школа экономики», Нижний Новгород, Россия

Ключевые слова: автоматическое создание языковых упражнений, автоматическая обработка текста, обучение иностранному языку с помощью компьютера, лексико-грамматические упражнения, анализ сложности текста, английский язык как иностранный

EXERCISE MAKER: AUTOMATIC LANGUAGE EXERCISE GENERATION

Malafeev A. Yu. (aumalafeev@hse.ru)

National Research University Higher School of Economics,
Nizhny Novgorod, Russia

Current trends in education, namely blended learning and computer-assisted language learning, underlie the growing interest to the task of automatically generating language exercises. Such automatic systems are especially in demand given the variability in language learning. Despite the abundance of resources for language learning, there is often a lack of specific exercises targeting a particular group of learners or ESP course. This paper gives an overview of a computer system called Exercise Maker that is aimed at flexible and versatile language exercise generation. The system supports seven exercise types, which can be generated from arbitrary passages written in English. Being able to tailor educational material to learners' interests is known to boost motivation in learners (Heilman et al., 2010). An important feature of the system is the automatic ranking of the source passages according to their complexity/readability. As shown by expert evaluation, the automatically generated exercises are of high quality: the gap precision is about 97–98%, while the overall exercise acceptance rate varies from 90% to 97.5%. Exercise Maker is freely available for educational and research purposes.

Key words: language exercise generation, natural language processing, computer-assisted language learning, lexico-grammatical exercises, text readability, English as a foreign language (EFL)

1. Introduction

Although commonly used, language tests and exercises are expensive to create manually. To address this issue, several systems for automatic language exercise generation have been developed in the past two decades. The exercises that can be automatically generated by these systems vary greatly in terms of the target aspects of vocabulary and grammar, supported languages, flexibility and effectiveness (see the next section for more details). From the pedagogic perspective, using automatically generated content seems especially relevant considering the modern educational trends, namely blended learning (Graham, 2006) and computer-assisted language learning (Levy, 1997).

In this paper, we give an overview of a computer program called "Exercise Maker" that we developed for automatic generation of lexical and grammatical exercises from arbitrary passages written in English. The program features seven different exercise types that can be customized to accommodate various learning needs. Importantly, Exercise Maker allows the user to adjust the difficulty: even if the same source passage is used, the exercises can be more challenging or less demanding. Currently, a Japanese version of the program is also under development.

The paper is structured as follows: Section 2 presents an overview of existing systems for automatic language exercise generation; Section 3 describes our solution; Section 4 presents some results of the evaluation of the effectiveness of the system; in the last section, we make conclusions and outline some possible ways of improving Exercise Maker.

2. Related Work

Some recent research has been conducted with a view to facilitating exercise creation. Among the more general solutions are multi-domain exercise or test generation systems, e.g. (Almeida et al., 2013; Mitkov et al., 2006; Sonntag, 2009) exercise generation systems usually work with answers of simple types (e.g. multiple-choice, Boolean, integer, or file comparison, as well as authoring tools, e.g. Hot Potatoes (http://hotpot.uvic.ca/), MaxAuthor (http://cali.arizona.edu/docs/wmaxa/) and others.

There are also several systems that are designed for generating exercises of one or more specific types to aid learners of the supported language(s). These are very different not only in the types of exercises they are able to generate, but also in terms of supported languages, type of input/output, external dependencies and whether the system is freely available. See Tab. 1 for details on these differences and a comparison of Exercise Maker with other systems. This is by no means a complete list of exercise generating systems, but rather some of the better-known and often cited ones. We believe that this is sufficient to demonstrate the prevalent trends in exercise generation and how Exercise Maker attempts to 'fill' some of the 'gaps'.

Tab. 1. Comparison of systems

No.	System	Languages	Input	Output	Exercise type(s)	External dependencies	Freely available
1	(Aldabe et al., 2006)	Basque	corpora	sentences	**fill-in-the-blank**, **word formation**, multiple choice, **error correction**	corpora; morpho-syntactic and syntactic parsers, phrase chunker	?
2	(Antonsen et al., 2013)	two Saami languages	lexicon and syntactic rules	sentences	**morphological transformation**	**none**	**yes**
3	(Bick, 2005)	**English** and 6 other	corpora	sentences	**open cloze**, **morphological transformation**	corpora	**yes**
4	(Brown et al., 2005)	**English**	words	questions	definition, synonym, antonym, hypernym, hyponym, and cloze questions (multiple choice or **wordbank**)	WordNet; external word frequency database	yes?
5	(Burstein and Marcu, 2005)	Arabic → **English**	corpora	sentences	translation	corpora; Arabic-to-English machine translation system	no
6	(Dickinson and Herring, 2008)	Russian	lexicon and syntactic rules	sentences	**morphological transformation**, **error correction**	none	no
7	(Gates, 2008)	**English**	texts from a corpus	questions	reading comprehension questions (factoid)	corpus; syntactic parser, lemmatizer, named entity extractor, semantic argument extractor, WordNet, parse tree transformer	no
8	(Goto et al., 2010)	**English**	**arbitrary texts**	questions	multiple choice	POS-tagger, web search	no
9	(Heilman and Eskenazi, 2007)	**English**	thesaurus	questions	finding related words	dependency parser, corpus	yes?
10	(Hoshino and Nakagawa, 2005)	**English**	**arbitrary texts**	sentences	multiple choice	WordNet	no
11	(Knoop and Wilske, 2013)	**Enslish**	**arbitrary texts**	sentences	multiple choice	WordNet	no
12	(Meurers et al., 2010)	**English**	arbitrary web-pages	**text**	**morphological transformation**, multiple choice, **open cloze**	external NLP framework; a separate external POS-tagger and constraint grammar rules; lexical database	yes
13	(Perez-Beltrachini et al., 2012)	French	lexicon and syntactic rules	sentences	shuffle questions, **open cloze**	grammar traverser	no?
14	(Sumita et al., 2005)	**English**	corpora	sentences	multiple choice	corpora; web search	no
15	Exercise Maker	**English**	**arbitrary texts**	**text**	fill in missing words (no blanks), **open cloze**, word formation, **wordbank**, **morphological transformation** (verb forms), text fragments, **error correction**	**none**	**yes**

To summarize, Exercise Maker is significantly different from most other systems: although it supports English only, indeed a very popular language, it generates exercises from arbitrary passages, which is a feature of only three other systems. Moreover, the output is also text, i.e. the exercises are not sets of separate, unrelated sentences, like in most other systems, but the same passages as input, with some modifications (e.g. gapped words, artificial 'errors', etc.). This 'context-rich' format is very similar to the one used in Cambridge English certificate exams, such as FCE, CAE, CPE, and BEC (e.g. see Cambridge English: Advanced Handbook for Teachers, 2012), which are very well-known and well-established English language tests (Chalhoub-Deville and Turner, 2000). In addition, the same format is used in the Russian State Exam (RSE) in English. But perhaps most importantly, being able to use any passages in English (rather than corpora or grammars and lexicons) means an opportunity to tailor educational material to learners' interests, which is known to boost learner motivation (Heilman et al., 2010).

Furthermore, some of the exercise types in Exercise Maker are not supported by other systems, namely filling in missing words (no gaps), word formation, and text fragments. Yet these types of exercises are commonly used in EFL, and some of them are included in FCE, CAE and CPE (word formation) and the RSE (word formation and text fragments). The difficulty of the exercises can be tweaked, which, although not shown in Tab. 1, is a very rare feature. Another important difference is that Exercise Maker is fully self-contained, which means that it can be more easily extended to resource-poor languages. Lastly, our system is freely available to anyone and, therefore, can be used not only for teaching and learning English, but also for research purposes, e.g. for comparison with other exercise-generating systems.

The next section will discuss the methods used for generating lexico-grammatical exercises in the Exercise Maker system.

3. Automatic Exercise Generation

Exercise Maker supports seven exercise types, which are listed in Tab. 2 with some additional information. This includes which exams, if any, use this type of task, as well as short descriptions and examples for each supported activity. The examples are generated by Exercise Maker using an input passage adapted from a Wikipedia article (http://en.wikipedia.org/wiki/Aron_Ralston).

Our system is implemented in Python and uses the standard libraries only. The generation method used is decision trees with manually written rules, although the exact algorithms vary depending on the exercise type. The rules often involve consulting a set of linguistic resources, specifically compiled by the author (manually and semi-automatically) for exercise generation. The linguistic resources are:

1. Two lists of 2274 and 10084 most common English word forms (including proper nouns), based on a free film-subtitle-based frequency list (https://invokeit.wordpress.com/frequency-word-lists/).
2. A list of 11805 word forms used in the word formation exercise heavily based on the BNC lists (http://simple.wiktionary.org/wiki/Wiktionary:BNC_spoken_freq).

3. A list of rules for making realistic spelling/lexical/grammar errors (795 words). The spelling part is based on the Wikipedia list of common misspellings (http://en.wikipedia.org/wiki/Wikipedia:Lists_of_common_misspellings), while the lexical and grammar error rules were compiled manually.
4. Three ordered lists of 139 words each for generating open cloze tests emulating specific Cambridge exam levels (FCE, CAE or CPE), based on an empirical study of the mentioned exams.
5. A list of 91 adverbs used in the verb forms exercise.
6. A list of 13540 verb forms and an additional short list of auxiliary forms, both used in the verb forms exercise. The lists were extracted from the Spelling Checker Oriented Word List (http://wordlist.sourceforge.net/).
7. A few manually written shorter lists of articles, conjunctions, prepositions, pronouns, etc.

Tab. 2. Exercise types supported by Exercise Maker

No.	Exercise	Exams	Description	Example(s)	Answer(s)
1	Word formation	FCE, CAE, CPE, RSE	Fill in blanks with derivatives of the words in parentheses.	...but the tools he had available were (6)_____ (sufficient) to do so.	insufficient
2	Error correction	BEC[1]	Correct spelling/lexical/grammar errors in the text.	Ralston had not informed nobody of his hiking plans <...> thus no one would searching for him <…> the dehydrated and deliriouse Ralston	had not informed anybody; no one would search/ be searching for him; delirious
3	Open cloze	FCE, CAE, CPE, BEC	Fill in blanks with suitable words (no candidate answers given). Sometimes, there are two or more correct answers.	When he ran (12)_____ of food and water on the fifth day…	out
4	Wordbank	none[2]	Fill in blanks with suitable words given a full list of answer choices (no distractors; each word is used only once).	(approximately, available, <…>, just, suspended) ...a (2)_____ boulder he was climbing down became dislodged…	suspended
5	Missing words (articles or prepositions)	none	Insert prepositions (another subtype: articles) where appropriate.	Ralston had not informed anybody his hiking plans, thus no one would be searching him.	Ralston had not informed anybody of his hiking plans, thus no one would be searching for him.
6	Text fragments	RSE	Insert missing text fragments (all answer options are listed).	After three days of trying to lift (6)_____, the dehydrated and delirious Ralston	d) and break the boulder
7	Verb forms	RSE	Use the appropriate verb form to fill each of the gaps.	While he (1)_____ (descend) a slot canyon, a suspended boulder…	was descending

[1] A somewhat similar task, but with only one error type—extra words.
[2] However, most exams use a somewhat similar test, the multiple choice.

It should be noted that the generation process does not consist in merely looking up words in the mentioned lists. For each type of exercise, there are rules that take into account such factors as capitalization, spelling features, punctuation, word length, distance to other gaps, word context, sentence boundaries, and others. Some rules may be quite complex. For example, dictionary look-ups do not suffice in the missing words (prepositions) exercise, because many words in English are ambiguous with respect to their part of speech. Thus, to determine that *to* is a preposition rather than a particle in a given context (both are common cases), the system checks if the next word is a determiner, or is capitalized, or contains a digit, or is the beginning of a new sentence, or it is longer than five characters and ends in *-ing*.

Another example of using rules beyond dictionary look-ups is making one common error in the error correction exercise. The error is misspelling adjectives ending in *-ous* and *-ful*, such as *furious* and *powerful*, to make these **furiouse* and **powerfull*. It would be difficult to list all possible adjectives that can be modified in this way, so the system uses the following simple rule: if the word is not all caps (avoids clashes with abbreviations) and ends in *-ous* or *-ful*, replace the ending with *-ouse* or *-full*, correspondingly.

These are merely some examples of the rules used in exercise generation; the size constraints do not permit listing all the rules.

The preprocessing step, performed once for each source passage, includes segmenting the input text into words, sentences and paragraphs, and analyzing the readability of the source. Text readability has a number of formal, quantitative characteristics, such as the average number of syllables in words and the average sentence length (Kincaid et al., 1975), or word frequency with respect to either a reference list (Chall, 1995) or corpus data (Stenner, 1996). Admittedly, these measures are error-prone and may be inaccurate at times, but they are still highly useful for approximating source text complexity. In Exercise Maker, the input passages are ranked according to their complexity. The latter is an important feature, as it helps the teacher to select materials appropriate for the ability level of particular learners, as, obviously, the readability of the source passage strongly correlates with the difficulty of the resulting exercises.

After experimenting with some variables, we chose two of them as the main proxy for text readability, namely the average sentence length and word frequency information. These two factors have traditionally been considered as the most closely correlated with text readability (Klare, 1968; Chall, 1995)3,15]]}},"label":"page"},{"id":315,"uris":["http://zotero.org/users/1547774/items/24IK5VWR"],"uri":["http://zotero.org/users/1547774/items/24IK5VWR"],"itemData":{"id":315,"type":"book","title":"Readability revisited: The new Dale-Chall readability formula","publisher":"Brookline Books Cambridge, MA","volume":"118","source":"Google Scholar","shortTitle":"Readability revisited","author":[{"family":"Chall","given":"Jeanne Sternlicht"}],"issued":{"date-parts":[["1995"]]}},"label":"page"}],"schema":"https://github.com/citation-style-language/schema/raw/master/csl-citation.json"} . While the first is very easy to calculate, the second may be approached in various ways. Similarly to (Chall, 1995), we use reference lists to approximate word frequency. Specifically, we have two reference lists of 2,274 and 10,084 most common English word forms, including proper nouns; these represent two levels of word frequency. If a word is in the first list, which means that it is very commonly used in English, it is also a member of the second, larger list.

Common, but less frequent words are those that are in the second list and not in the first. If a word is not a member of either of the lists, it is considered an 'unknown' word. Readability is thus determined by the proportion of 'unknown' words in a text, the proportion of words that are not in the first list, and the average sentence length. These three factors have equal weight in our simple complexity model.

Importantly, Exercise Maker goes beyond readability in adjusting the difficulty of exercises. For almost every type of exercise (except missing words), the system generates several subtypes with varying settings that affect the difficulty. These settings are:

- number of gaps in the exercise;
- target language material, i.e. the words in the text that are gapped (for the open cloze and verb forms exercises);
- length of the gaps (for the fragments exercise).

Regarding target words, it might be necessary to clarify that the open cloze exercises are based on five different lists of target word forms, which is aimed at generating exercises of varying difficulty and at emulating specific Cambridge exam task types: FCE, CAE and CPE. As for the verb form exercises, these come in two varieties: gapping 'simple', one-word verb forms and more complex, multiword verb forms.

In the next section, we describe an experiment conducted to evaluate the performance of our system.

4. Evaluation

Although TEFL experts in several educational establishments have successfully used our system, it is necessary to present here a formal evaluation of Exercise Maker. Earlier, we evaluated specific exercise types such as the open cloze (Malafeev, 2014) and got interesting results. In particular, two groups of TEFL experts (17 and 16 people) found it considerably difficult to tell the difference between activities generated by Exercise Maker and tests authored by Cambridge professionals.

For this publication, we conducted a specific evaluation session covering all seven exercise types. Two independent TEFL experts, both non-native speakers of English, who had not taken any part in developing Exercise Maker, participated in the evaluation. We downloaded five abridged and simplified news articles from a popular website for EFL teachers and learners, breakingnewsenglish.com (see Tab. 3). We had not read or otherwise used these articles prior to the evaluation experiment.

Tab. 3. News articles used for evaluation

No.	Title	Date	Word count	Readability[1]
1	Japanese government to play matchmaker	17th March, 2015	243	10.7
2	BBC Top Gear star punches producer	14th March 2015	237	7.1
3	Sportswear maker accused of sexism	11th March, 2015	231	12.1
4	China tops US at box office for first time	5th March, 2015	233	7.7
5	Cut music to an hour a day	2nd March, 2015	248	12.2

[1] For readability, we used the Automated Readability Index (Kincaid et al., 1975), calculated using the Edit Central online service (http://www.editcentral.com/gwt1/EditCentral.html).

The five articles were used as input to generate 40 exercises with our system, eight from each text. Although the number of exercise types supported by the system is seven, we chose to use two different subtypes of the missing words exercise, namely articles and prepositions.

The experts had to perform two kinds of assessment:

- evaluate all gaps in all exercises and determine which of the gaps are valid, i.e. potentially useful in teaching or testing, and which are not (evaluating precision only is a widely accepted practice in automatically-generated exercise evaluation);
- assign to each exercise an overall score from 1 to 4, meaning:
  1 – the exercise cannot be used;
  2 – the exercise can be used only after making substantial alterations;
  3 – the exercise can be used, but it requires some minor alterations;
  4 – the exercise can be used as is.

In the latter form of assessment, the scores of 3 or 4 would mean that the exercise is 'acceptable', and the lower scores would mean that it is not.

The experts were supplied with detailed instructions written in Russian on the evaluation procedure. The evaluation took about three hours (expert 1) and five hours (expert 2). We believe that, given that both the articles and our system are freely available for download, and the assessment guidelines are available on request, our experiment can be easily reproduced, although, of course, with different experts. The results of the evaluation are presented in Tab. 4, Tab. 5 and Tab. 6.

Tab. 4. Evaluation results, validity of gaps (precision)

	Evaluation	Total gaps	Expert 1, n	Expert 1	Expert 2, n	Expert 2
Exercises	Articles	110	109	99,09%	110	100,00%
	Derivatives	60	52	86,67%	57	95,00%
	Errors	131	131	100,00%	129	98,47%
	Fragments	30	30	100,00%	30	100,00%
	Open cloze	144	142	98,61%	144	100,00%
	Prepositions	137	126	91,97%	131	95,62%
	Verb forms	73	73	100,00%	69	94,52%
	Wordbank	100	97	97,00%	100	100,00%
Texts	1	153	145	94,77%	151	98,69%
	2	171	165	96,49%	168	98,25%
	3	153	147	96,08%	146	95,42%
	4	144	142	98,61%	142	98,61%
	5	164	161	98,17%	163	99,39%
Total	micro	785	760	96,82%	770	98,09%
	macro, exercises			96,67%		97,95%
	macro, texts			96,92%		97,80%

Tab. 5. Evaluation results, accepted exercises

	Evaluation	Total exercises	Accepted by expert 1		Accepted by expert 2	
			n	%	n	%
Exercises	Articles	5	5	100,00%	5	100,00%
	Derivatives	5	4	80,00%	4	80,00%
	Errors	5	5	100,00%	5	100,00%
	Fragments	5	4	80,00%	5	100,00%
	Open cloze	5	5	100,00%	5	100,00%
	Prepositions	5	3	60,00%	5	100,00%
	Verb forms	5	5	100,00%	5	100,00%
	Wordbank	5	5	100,00%	5	100,00%
Texts	1	8	6	75,00%	8	100,00%
	2	8	7	87,50%	8	100,00%
	3	8	8	100,00%	7	87,50%
	4	8	7	87,50%	8	100,00%
	5	8	8	100,00%	8	100,00%
Total		40	36	**90,00%**	39	**97,50%**

Tab. 6. Scores assigned by the experts

Score	Expert 1		Expert 2	
	n	%	n	%
1	0	0,00%	0	0,00%
2	4	10,00%	1	2,50%
3	22	55,00%	8	20,00%
4	14	35,00%	31	77,50%

As can be seen from the tables, the gap precision is about 97–98%, which is very high. The acceptance rate varies significantly, from 90% (expert 1) to 97.5% (expert 2). This difference can probably be explained by the fact that the borderline between "substantial alterations" and "minor alterations" is not well-defined and depends on the subjective judgment, even with assessment guidelines. Besides, as commented by the first expert, while individual gaps seemed valid, the combination of these did not always produce a good exercise. Indeed, different TEFL professionals might have varying opinions about what exactly constitutes a good language exercise. Still, we believe that even the lower, 90% acceptance rate is a very good result for automatic language exercise generation.

The next section will draw conclusions and outline some possible directions for future work.

5. Conclusion

This paper presents an overview of our language exercise generation system, Exercise Maker. With it, a variety of lexical and grammatical exercises can be automatically generated from arbitrary passages written in English. The source passages are ranked according to their readability to help the user choose appropriate material. The seven types of supported exercises can be further customized to accommodate various learning needs. Besides, Exercise Maker allows the user to adjust the difficulty, even if the same source passage is used. As shown in the evaluation section, the exercises generated are perceived by TEFL experts as quite useful.

The most promising directions of future work are the following:

- support for other languages;
- new exercise types, such as multiple choice;
- further improving exercise quality, possibly with statistical methods and machine learning.

Acknowledgements

The author thanks the three anonymous reviewers for valuable comments and suggestions.

References

1. *Aldabe, I., Lacalle, M. L. de, Maritxalar, M., Martinez, E., and Uria, L.* (2006). ArikIturri: An Automatic Question Generator Based on Corpora and NLP Techniques. In Intelligent Tutoring Systems, M. Ikeda, K. D. Ashley, and T.-W. Chan, eds. (Springer Berlin Heidelberg), pp. 584–594.
2. *Almeida, J. J., Araujo, I., Brito, I., Carvalho, N., Machado, G. J., Pereira, R. M. S., and Smirnov, G.* (2013). PASSAROLA: High-Order Exercise Generation System. In 2013 8th Iberian Conference on Information Systems and Technologies (CISTI), pp. 1–5.
3. *Antonsen, L., Johnson, R., Trondsterud, T., and Uibo, H.* (2013). Generating Modular Grammar Exercises with Finite-State Transducers. Proceedings of the Second Workshop on NLP for Computer-Assisted Language Learning at NODALIDA 2013 27–38.
4. *Bick, E.* (2005). Live Use of Corpus Data and Corpus Annotation Tools in CALL: Some New Developments in VISL. Nordic Language Technology, AArbog for Nordisk Sprogteknologisk Forskningsprogram 171–185.
5. *Brown, J. C., Frishkoff, G. A., and Eskenazi, M.* (2005). Automatic Question Generation for Vocabulary Assessment. In Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, (Stroudsburg, PA, USA: Association for Computational Linguistics), pp. 819–826.

6. *Burstein, J., and Marcu, D.* (2005). Translation Exercise Assistant: Automated Generation of Translation Exercises for native-Arabic Speakers Learning English. In Proceedings of HLT/EMNLP on Interactive Demonstrations, (Stroudsburg, PA, USA: Association for Computational Linguistics), pp. 16–17.
7. *Chalhoub-Deville, M., and Turner, C. E.* (2000). What to Look for in ESL Admission Tests: Cambridge Certificate Exams, IELTS, and TOEFL. System 28, 523–539.
8. *Chall, J. S.* (1995). Readability revisited: The new Dale-Chall readability formula (Brookline Books Cambridge, MA).
9. *Dickinson, M., and Herring, J.* (2008). Developing Online ICALL Exercises for Russian. In Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications, (Stroudsburg, PA, USA: Association for Computational Linguistics), pp. 1–9.
10. *Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., and Willingham, D. T.* (2013). Improving Students' Learning With Effective Learning Techniques Promising Directions From Cognitive and Educational Psychology. Psychological Science in the Public Interest 14, 4–58.
11. *Gates, D. M.* (2008). Automatically Generating Reading Comprehension Look-Back Strategy: Questions from Expository Texts (Pittsburgh: Carnegie Mellon University).
12. *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: An International Journal (KM&EL) 2, 210–224.
13. *Graham, C. R.* (2006). Blended learning systems. CJ Bonk & CR Graham, The Handbook of Blended Learning: Global Perspectives, Local Designs. Pfeiffer.
14. *Heilman, M., and Eskenazi, M.* (2007). Application of Automatic Thesaurus Extraction for Computer Generation of Vocabulary Questions. In Proceedings of the SLaTE Workshop on Speech and Language Technology in Education, pp. 65–68.
15. *Heilman, M., Collins-Thompson, K., Callan, J., Eskenazi, M., Juffs, A., and Wilson, L.* (2010). Personalization of Reading Passages Improves Vocabulary Acquisition. International Journal of Artificial Intelligence in Education 20, 73–98.
16. *Hoshino, A., and Nakagawa, H.* (2005). WebExperimenter for Multiple-choice Question Generation. In Proceedings of HLT/EMNLP on Interactive Demonstrations, (Stroudsburg, PA, USA: Association for Computational Linguistics), pp. 18–19.
17. *Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., and Chissom, B. S.* (1975). Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel (DTIC Document).
18. *Klare, G. R.* (1968). The Role of Word Frequency in Readability. Elementary English 12–22.
19. *Knoop, S., and Wilske, S.* (2013). WordGap—Automatic Generation of Gap-Filling Vocabulary Exercises for Mobile Learning. Proceedings of the Second Workshop on NLP for Computer-Assisted Language Learning at NODALIDA 2013 39–47.
20. *Levy, M.* (1997). Computer-Assisted Language Learning: Context and Conceptualization. (ERIC).

21. *Malafeev, A.* (2014). Automatic Generation of Text-Based Open Cloze Exercises. In Analysis of Images, Social Networks and Texts, D. I. Ignatov, M. Y. Khachay, A. Panchenko, N. Konstantinova, and R. E. Yavorsky, eds. (Cham: Springer International Publishing), pp. 140–151.
22. *Meurers, D., Ziai, R., Amaral, L., Boyd, A., Dimitrov, A., Metcalf, V., and Ott, N.* (2010). Enhancing Authentic Web Pages for Language Learners. In Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications, (Stroudsburg, PA, USA: Association for Computational Linguistics), pp. 10–18.
23. *Mitkov, R., An Ha, L., and Karamanis, N.* (2006). A Computer-Aided Environment for Generating Multiple-Choice Test Items. Natural Language Engineering 12, 177–194.
24. *Perez-Beltrachini, L., Gardent, C., and Kruszewski, G.* (2012). Generating Grammar Exercises. In Proceedings of the Seventh Workshop on Building Educational Applications Using NLP, (Stroudsburg, PA, USA: Association for Computational Linguistics), pp. 147–156.
25. *Sonntag, M.* (2009). Exercise Generation by Group Models for Autonomous Web-Based Learning. In 35th Euromicro Conference on Software Engineering and Advanced Applications, 2009. SEAA '09, pp. 57–63.
26. *Stenner, A. J.* (1996). Measuring Reading Comprehension with the Lexile Framework. Fourth North American Conference on Adolescent/Adult Literacy, Washington DC.
27. *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. In Proceedings of the Second Workshop on Building Educational Applications Using NLP, (Stroudsburg, PA, USA: Association for Computational Linguistics), pp. 61–68.
28. *Cambridge English* (2012): Advanced Handbook for Teachers (Cambridge ESOL).