Measuring the Power of Learning.®

Research Report
ETS RR–17-52

Analyzing Item Generation with Natural Language Processing Tools for the *TOEIC*® Listening Test

Su-Youn Yoon
Chong Min Lee
Patrick Houghton
Melissa Lopez
Jennifer Sakano
Anastassia Loukina
Bob Krovetz
Chi Lu
Nitin Madnani

December 2017

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Analyzing Item Generation with Natural Language Processing Tools for the *TOEIC*® Listening Test

Su-Youn Yoon, Chong Min Lee, Patrick Houghton, Melissa Lopez, Jennifer Sakano, Anastassia Loukina, Bob Krovetz, Chi Lu, & Nitin Madnani

Educational Testing Service, Princeton, NJ

In this study, we developed assistive tools and resources to support *TOEIC*® Listening test item generation. There has recently been an increased need for a large pool of items for these tests. This need has, in turn, inspired efforts to increase the efficiency of item generation while maintaining the quality of the created items. We aimed to address this challenge by creating a set of automated tools and resources that support item generation: an automated system that retrieves appropriate real-world videos, a list of vocabulary tagged with established difficulty levels, and a tool that suggests words and phrases that are similar in distribution to a given word (word similarity tool). These tools and resources were designed to help item writers by providing initial ideas, authentic language, and support for adjusting the variety and complexity of vocabulary in their items. To evaluate the impact of these resources on the efficiency of item generation, seven item writers created listening items using our tools. All tools were considered useful, and the word similarity tool in particular was rated the most useful. The tools are currently applied to English item generation for the TOEIC Listening test, but the method is generic and applicable to other languages.

Keywords TOEIC Listening test; item generation; automated assistive tools for item generation; natural language processing; survey

doi:10.1002/ets2.12183

Recent developments in natural language processing (NLP) technology and massive online resources have substantially changed the environment of language learning. Online materials are useful sources of authentic situations and language use, and they have been frequently used in generating vocabulary lists (Capel, 2010; Coxhead, 2000; Fuentes, 2002) and in examples of collocation expressions (Chen, Huang, Chang, & Liou, 2015; Liou et al., 2013).

Another frequent use of online resources and computerized corpora is the development of listening and reading materials. Several studies have explored the application of NLP technologies to the selection of appropriate reading or listening materials for students who have English as a second language (ESL), with most studies focused on evaluating the difficulty of materials. Graesser, McNamara, Louwerse, and Cai (2004) and Sheehan, Kostin, Napolitano, and Flor (2014) developed automated systems to provide the overall difficulty score of written text in English. One of the primary goals of these systems is to provide native and ESL students with reading or listening materials according to their grade or language proficiency level. These studies focused on estimating difficulty for the given texts.

In order to create high-quality reading or listening materials from online resources, aspects other than difficulty also need to be considered. Heitler (2005) developed a manual on how to prepare classroom materials from online resources and provided useful strategies such as adjusting text length, replacing vocabulary, simplifying syntactic structure, and resolving proper names and abbreviations. To reduce this manual effort, an initial selection of appropriate materials that can be quickly and easily adapted into learning materials is necessary. However, few studies have discussed the characteristics of such appropriate materials. Furthermore, previous studies have mostly focused on generating learning materials (e.g., classroom materials) and have not discussed characteristics of appropriate materials for language assessments. As mentioned in Hoshino and Nakagawa (2007), automated material selection is more difficult to apply to language assessments than learning materials because the former is subject to greater strictures on language variety, type, and difficulty. There may be additional requirements that the online resources must meet in order to be considered appropriate assessment materials, and these requirements increase the difficulty of fully automated material selection.

Corresponding author: S.-Y. Yoon, E-mail: syoon@ets.org

In this study, we developed assistive tools based on NLP technology and online resources to support listening item generation for *TOEIC*® Listening, a large-scale international English proficiency test. In contrast to previous studies, which focused on the automated generation of limited item types such as a cloze test for vocabulary and prepositions (e.g., Heilman & Smith, 2010; Huang, Chen, & Sun, 2012; Huang, Tseng, Sun, & Chen, 2014), our tools support diverse tasks for a multitude of different item types.

We developed three tools: an automated system that retrieves appropriate real-world videos, a list of vocabulary associated with difficulty levels, and a tool that suggests words that occur frequently in similar contexts. These tools were expected to improve the quality of items by increasing the diversity and authenticity of contexts and vocabulary, which would also increase the efficiency of item generation because diversity and authenticity prevent overlap among items and reduce the amount of revision as a result.

This study addresses following points:

- We provide a discussion about the characteristics of appropriate materials for listening items based on an annotation study with two expert language test developers.
- We provide three tools that support the main tasks related to listening item generation: passage generation, adjustment of the word difficulty used in items, and distractor generation.
- We examine the usefulness of these tools through a small-scale item generation study.

Assistive Tools Used in this Study

In this study, we classified listening item generation process into three stages and developed a tool to support each stage as follows:

- Brainstorming and idea generation: seed video retrieval system
- Distractor generation: word similarity tool
- Revision and adjustment of the created item: vocabulary list

Word Similarity Tool

We used a tool to identify words that convey similar meanings (e.g., student and learner) or related meanings (e.g., student and school) developed by Heilman and Madnani (2012). Using NLP techniques, researchers employed empirical approaches to assess lexical associations. Based on the intuition that words with similar contextual distribution (i.e., the linguistic contexts that they appear in) will have similar or related meanings, they calculated distributional similarities among words from large text corpora. Following this line of research, we first estimated distributional similarities among words based on Dekang Lin's Distributional Thesaurus and stored them in a large database. We provided a Web-based user interface, and it returned the 10 most similar words, based on similarity score, given the query word provided by the item writers.

Vocabulary List

We created a vocabulary list by combining the following three vocabulary lists:

- New General Service List (NGSL): A word list designed for general service purposes. The list is composed of the 2,800 most frequently occurring words extracted from a subset of the Cambridge English corpus, which includes approximately 270 million words.
- Lemmatized British National Corpus (BNC) frequency list: A word list including the top 6,318 most frequently occurring words from the BNC
- Corpus-based list: A word list including the top 7,699 most frequently occurring words from an English Gigaword corpus

The words from these three lists were classified into four groups. First, we made a separate category for 368 function words such as articles, prepositions, and pronouns. Next, we classified the remaining vocabulary into three tiers: Tier 1 for basic vocabulary, Tier 2 for intermediate level vocabulary, and Tier 3 for topic-specific vocabulary. As the tier increases, the difficulty of the vocabulary also increases. The difficulty level (the tier the word belongs to) was determined based on the source, rather than its frequency in the specific corpus. A total of 2,551 words in the NGSL list excluding function words were assigned to Tier 1; 1,712 words in the BNC but not in the NGSL and function words were assigned to Tier 2; and 3,478 words in the mixed corpora-based list but not in NGSL, BNC, or function words were assigned to Tier 3. The final list was composed of 8,109 unique words.

We also provided a separate vocabulary list created using a large pool of listening items. The item corpus was composed of 19,460 listening items extracted the TOEIC Listening test. The list included a total of 3,503 unique words, their tier information when available, and the number of items that include this word.

Automated Seed Video Retrieval System

Good items make use of authentic language used in varied situations. Writing a listening item that provides an appropriate level of difficulty, reflecting authentic language use while avoiding duplicates, is a difficult task. Writing such an item about unfamiliar topics is an especially challenging task for item writers and it requires a substantial amount of research to find initial ideas.

The most frequently used approach to develop items in unfamiliar contexts is searching Web resources. However, finding the appropriate materials with unguided searching is not an easy task, and item writers tend to spend time reviewing useless Web resources retrieved from search engines. Results from Web searches usually contain a large amount of irrelevant material to sift through, including redundant or unrelated content as well as content inappropriate for language assessments.

In order to address this issue, we designed an assistive tool, called the Seed Video Retrieval System, for test item writers. This tool provides Web resources that have a greater likelihood of being useful for writing test items. When more helpful Web resources are available, item writers can reduce time wasted finding resources from retrieved search results. This system is designed to retrieve only YouTube videos that meet certain constraints, when users enter search keywords.

Although text resources are also available as Web resources, we decided to focus on YouTube videos due to a few advantages of videos over text resources. After an initial attempt to use Web pages as resources for item writers, we observed challenges and drawbacks in extracted text resources. Item writers wanted a small number of concise data sources relevant to their keywords. They also wanted data containing contexts with enough development to allow them to understand the content. The power of videos in conveying content is well expressed in an idiom: a picture is worth a thousand words. Visual images in videos can provide more information to the viewer than words in texts. So, videos can be more concise while providing as much or more information. For example, it can be easier to figure out which vocabulary words need to be used in which situation when an item writer watches a video. The images in a video inherently contain lots of contextual information on places, tools, roles, and so forth. Furthermore, extracted text data usually contained too much text to read and texts on topics outside item writers' fields of expertise, which required further research to understand. Moreover, Web pages usually contained redundant data such as HTML tags and content irrelevant to search keywords. It was a technical challenge to automatically remove the redundant or irrelevant data from the set of retrieved pages, in order to provide a useful tool.

During our initial exploration of YouTube videos as a resource, we discovered some challenges for item writers who might seek to use them. Some videos were too long, too difficult to understand, or too incoherent to make items. We will further discuss the characteristics of appropriate videos in the Participants section. Based on a qualitative analysis using a subset of data, we found that a higher percentage of videos with manual transcriptions contained coherent content and better audio quality. Manual transcription means that the video's owners provided transcripts of speech in the videos when they were uploaded. The existence of manual transcriptions could be indirect evidence that uploaders paid more attention to the quality of the videos and that they also considered their audiences. As a result, a higher percentage of such videos were appropriate for item generation than was the case for videos lacking these manual transcriptions. Based on these findings, we developed an automated system using the YouTube API with refined search conditions. The system retrieved videos with a manual transcription shorter than 4 minutes in length.

When we tested our video retrieval system, we found that the search skills differed greatly across the individual item writers, and the usefulness of the tool also substantially varied depending on their skills. Therefore, instead of providing the tool itself, we created a set of key words by concatenating topic and genre words provided by expert item writers. We selected four topics and collected a total of 664 videos. For each video, the title, the key words used in the video search, and the link to video were provided in an Excel spreadsheet. The quality of this video data collection is analyzed in the Participants section.

User Study

In order to investigate the usefulness of the assistive tools, we conducted a small-scale pilot study. The participants took part in an 8-week item-writing program, and during the program they were asked to use the tools described above to assist them in creating items. At the end of the program, the participants completed a survey and answered questions during a follow-up interview about the usefulness of these tools.

Participants

Applicants filled out a form where they created several types of common listening items. These were scored blindly by experienced item writers, organizers of the 8-week item-writing program, without any personal information about the applicant. The selected item writers consisted of six women and one man. Their educational backgrounds included undergraduate students with different majors (e.g., French, history, education, and journalism), a university professor teaching English to speakers of other languages, and a public school teacher of bilingual education. Two of the participants have substantial experience in item generation and participated in the same item-writing program for 3 years. The other five item writers were first-time participants.

Tasks

The participants were asked to create the following three types of TOEIC Listening items:

- Type 1: The test taker is presented with a picture and four recorded statements and asked to select the statement that best describes the picture.
- Type 2: The test taker listens to a conversation between two speakers and answers a series of written questions about the content of the conversation.
- Type 3: The test taker listens to a recording of a single speaker (e.g., announcement or advertisement) and answers a series of written multiple-choice questions about the content of the recording.

The tools were introduced to participants in the second week of the program. We provided a 30-minute presentation and question-and-answer session, as well as written manuals. Both the seed videos and vocabulary list were presented as spreadsheets, and the word similarity tool was presented as a website. All participants were requested to use the tools during first 2 weeks of the test period. After this initial test period, the use of tools was optional, but all participants used at least one tool throughout the entire program. During the 8-week program, each participant created 18 Type 1, 40 Type 2, and 40 Type 3 items, on average.

We asked participants about their usage of the tools using a survey and interview on the last day of the program. The survey was composed of two questions about the participants' background (experience in item generation) and 21 questions about their experience with the tools, divided over four sections in the survey: frequency of use, perceived usefulness, method of use, and future improvements. Multiple-choice questions were used for the frequency of use section. For perceived usefulness section, we used 12 Likert-type questions (four questions for each tool). Higher point responses indicated a higher degree of usefulness in item generation. Finally, open-ended questions were used for both the method of use and future improvement sections in the survey.

There were follow-up interviews after the survey responses were collected. Participants' survey responses were reviewed before the interviews, and two researchers in this study asked questions to understand survey responses further. Participants were asked to clarify why they did or did not use particular tools and how the tools were used in the item writing process and to expand on some of the shorter responses. In this way, we were able to pinpoint the ways in which the tools were successful and the aspects we could focus on improving. The interviews also allowed some context in which to evaluate the multiple-choice and Likert responses qualitatively.

In the next section in this report, we provide some insight into the participants' evaluation of the usefulness of the tools. Therefore, we focus on frequency of use, perceived usefulness, and method of use.

Table 1 Distribution of Annotations

Question	Annotator 1			Annotator 2		
	Yes	Maybe	No	Yes	Maybe	No
Seed video	286 (43%)	210 (32%)	168 (25%)	397 (60%)	209 (31%)	58 (9%)
New context	428 (64%)	164 (25%)	72 (11%)	509 (77%)	147 (22%)	8 (1%)
Sufficient info	283 (43%)	199 (30%)	182 (27%)	263 (40%)	258 (39%)	143 (22%)
Appropriate content	389 (59%)	161 (24%)	114 (17%)	488 (73%)	152 (23%)	23 (4%)
Formal language	597 (90%)	44 (7%)	23 (3%)	546 (82%)	68 (10%)	50 (8%)
Vocabulary difficulty	531 (80%)	73 (11%)	60 (9%)	511 (77%)	77 (12%)	76 (11%)

Results

To What Extent Are Automatically Retrieved Resources Appropriate for Item Generation?

In order to evaluate the quality of videos retrieved by the automated seed video retrieval system, two experienced item writers were recruited to rate 664 videos. First, they were asked to rate the holistic quality of each video with regard to its appropriateness as a seed video (Is the video helpful in item writing?). In addition, they answered the following five subquestions:

- Does the video contain a new context? (new context)
- Does the video contain sufficient information to understand it? (sufficient info)
- Is the content of the video appropriate for the test? (appropriate content)
- Does the video provide good examples of formal language? (formal language)
- Is the video generally appropriate in terms of vocabulary difficulty? (vocabulary difficulty)

The first question in the list above is about whether a retrieved video contains a new context that reflects contemporary language expressions and situations that have not been frequently used in existing items. The second question is about whether an annotator understands a given video without referring to other resources. The third question is about whether the content of a video could be used in test items. The fourth question serves to help figure out if the video contains words of a level of formality that is useful for test item writing. The fifth question is designed to explore the influence that the vocabulary difficulty of a video has on the usefulness of that video.

For each question, annotators were asked to choose one answer: yes, maybe, or no. *Yes* means that a video is highly likely to be qualified for the stated characteristic, *no* means that a video is highly unlikely to be qualified for the stated characteristic, and *maybe* means that a video is likely to be somewhat qualified for the stated characteristic.

Table 1 shows the distributions of annotators' answers on the questions. In addition, each cell contains a count and its ratio (a count of yes, maybe, or no divided by the count of all videos).

Annotator 1 and 2 considered 43% and 60%, respectively, of 664 videos to be appropriate seed videos that could be helpful in writing test items. The number of videos for which both annotators answered yes on the main question about appropriateness as a seed video was 243 (36.6%). The number of videos for which at least one annotator marked yes was 440 (66.3%). So, depending on item writers' needs, over half of the retrieved videos could be helpful in writing test items. In order to calculate the interannotator agreement, we converted ratings into a numeric scale: 1 for yes, 2 for maybe, and 3 for no. The quadratic weighted kappa on the main question was 0.51.

Both annotators thought that most videos (from 60% to 90%) met criteria for new context, appropriate content, formal language, and vocabulary difficulty; however, the proportion of videos that contained sufficient information was substantially lower (ranging from 40% to 43%). A possible reason that the majority of retrieved videos could meet the prescribed criteria was the search conditions we adopted. We only selected videos with manual captions and these results were in line with our expectations.

As an initial effort to develop an automated classifier that predicts the holistic quality of seed videos, we investigated to what extent the manual annotations of the five subquestions could accurately categorize the retrieved videos into *appropriate*, *maybe*, or *inappropriate* seed videos. We converted yes, maybe, and no answers into 1, 2, and 3, respectively, and then trained multiple linear regression models with the seed video question as a dependent variable and the five subquestions as independent variables. We used all 664 videos for the model building and reported model fits in the training

Table 2 Regression Analysis Using Annotated Data

Size of combination	Annotator 1 Features	R^2	Adjusted R^2	Annotator 2 Features	R^2	Adjusted R^2
1	New context	0.753	0.752	Appropriate content	0.562	0.561
2	New context + sufficient info	0.866	0.865	Appropriate content + new context	0.647	0.647
3	New context + sufficient info + appropriate content	0.880	0.880	Appropriate content + new context + sufficient info	0.700	0.700
4	New context + sufficient info + appropriate content + formal Language	0.884	0.883	Appropriate content + new context + sufficient info + formal language	0.717	0.717

Table 3 Frequency of Use for Each Tool

Tool	Never	Once	Once to a few times per week	Daily	Multiple times a day	Total
Word similarity tool	1	1	1	1	3	7
Vocabulary list	1	2	4	0	0	7
Seed videos	0	1	6	0	0	7

data. We tried all combinations of the five subquestions and reported the best performers for each size of combination in terms of the coefficient of determination (R-squared, R^2). We excluded vocabulary difficulty because including it in the model did not result in significant improvement in R^2 score. All of the regressions in Table 2 were significant at p-value <.001.

Table 2 shows which combinations of sub questions lead to improvements of both R^2 and adjusted R^2 values. For example, when only one factor is considered, new context and appropriate content were the best factors for Annotators 1 and 2, respectively. The best adjusted R^2 scores for Annotators 1 and 2 (.883 and .717, respectively) were achieved using the combination of new context, sufficient info, appropriate content, and formal language. Previous studies about automated listening and reading material selection mostly focused on difficulty, and other dimensions such as topics and contents have been neglected. This analysis shows the importance of these dimensions. In particular, for assistive item generation, they are the most important factors in determining the appropriateness of the materials.

Can Assistive Tools Improve the Item Generation Process?

Survey Questions 1, 2, and 3 from the pilot test of the tools solicited information about the frequency of use for each tool. All participants used at least one tool, once to a few times per week. Frequency of use for each tool is presented in Table 3. In general, the word similarity tool was the most frequently used among the three tools, and four participants (57%) used it more than once a day. It was followed in popularity by the seed video list and the vocabulary list.

Next, we asked the participants about how they used each tool during item generation. The participants provided a short description, and we got further detailed explanations during the follow-up interviews. The word similarity tool was used to find similar words (much like a thesaurus) and avoid repetition of words both in stimuli and items. One participant specifically mentioned it was used for distractor generation. The vocabulary list was used to adjust the difficulty of vocabulary in both stimuli and items. The tool was used both in addition of words (low or medium frequency words) and removal of words (high frequency words, removed to avoid repetition). One participant used it to create a list of context ideas by using a random word and phrase generator that was provided with the tool. The seed video tool was used primarily for idea generation of Item Types 2 and 3. In addition, three participants used videos to extract authentic language expressions and terms for specific fields.

Table 4 Perceived Usefulness of Each Tool

Tool		Strongly disagree	Somewhat disagree	Somewhat agree	Strongly agree	Not applicable
Word similarity tool	No. of responses	1	0	9	12	6
	%	4	0	32	43	21
Vocabulary list	No. of responses	0	3	9	2	14
	%	0	11	32	7	50
Seed videos	No. of responses	1	13	8	5	1
	%	4	46	29	18	4

Twelve survey questions solicited the perceived usefulness for each tool. We asked the following four question types:

- Speed of item generation: Using the tool enabled me to write items more quickly.
- Quality of created item: Using the tool improved the feedback I received for quality of my items.
- Diversity in context and vocabulary (subquestion for quality): Using the tool made it easier to create a larger variety of items (with regard to contexts, difficulty, etc.).
- Overall usefulness: I found the tool useful in my job.

Each question was a 4-point Likert scale, where 1 indicated strong disagreement and 4 indicated strong agreement. In addition, the participants could select not applicable if, for instance, they had not used a particular tool beyond the initial requested period or did not effectively use the tools in generating any items. Indeed, some participants found some of the tools to be more time-consuming than useful.

We investigated the usefulness of each tool, and the four questions for each tool were combined into a single composite score during analysis. A total of 28 responses (7 participants multiplied by 4 questions) were available for each tool. Table 4 summarizes the results.

We found the most positive responses for the word similarity tool. For the general usefulness question (Q4), six participants agreed or strongly agreed that the tool was useful in item generation, with four participants indicating strong agreement.

The vocabulary list was less frequently used than the other tools, and the participants chose not applicable for 50% of responses across four questions. When we excluded not applicable responses, participants provided substantially more positive reactions than negative reactions. For the general usefulness question (Q4), four participants (75% after excluding participants who chose not applicable) agreed or strongly agreed that the tool was useful in item generation. Thus, we can see that the tool was useful for the smaller group of participants who actually used it.

The seed video list was the most widely used of the tools, and the ratings varied across different question types. For the general usefulness question (Q4), four participants agreed or strongly agreed that the tool was useful in item generation, and the positive response was slightly more frequent than negative responses (three participants disagreed). The tool received the most positive evaluation for diversity of context and vocabulary (Q3), and five participants agreed or strongly agreed that the tool increased the variety in created items. The tool received the least positive evaluation for speed of item generation (Q1), with five participants disagreeing that the tool increased the speed of item generation. The tool was favored by one of the experienced item writers, who strongly agreed that the tool was useful in item generation. She pointed out that the seed video list may be more useful for experienced item writers who may have exhausted ideas for new items. The vocabulary tools, she felt, may be useful for novice item writers who are not yet familiar with the tasks and the kinds of vocabulary that are appropriate for potential test takers. This suggests potential differences in the usefulness of tools between experienced item writers and new item writers.

In an additional analysis, we converted each option to a numeric value and calculated the mean of Likert-scale scores for each tool. Strongly disagree, disagree, somewhat agree, and strongly agree were mapped into 1, 2, 3, and 4, respectively, and not applicable was excluded from analysis. The mean scores for the word similarity tool, vocabulary list, and seed video list were 3.45, 2.93, and 2.63, respectively, on a 4-point scale.

Finally, participants provided comments about how to improve tools. Many comments were related to the organization and presentation of the seed video collection. Because the participants were not assigned to create items on a specific topic, we initially hypothesized that participants may use any video if it included appropriate materials for the target language proficiency test. However, in reality, participants first made a decision about a narrow topic of the item and started searching videos relevant to the specific topic. As a result, an efficient interface to help search within the video data collection was required. Here are some detailed comments:

- Descriptions: In addition to the YouTube video title, the participants requested a short summary for each video.
- Content overlap: The video collection included multiple videos that were not identical but similar in content. The participants suggested removing videos with similar content to reduce the overlap.

Based on these comments, we are currently improving the automated seed video retrieval system. First, we will provide the category and video uploader information for each video, in order to improve the descriptions of videos. To reduce the overlapping content, we set a limit on the number of videos from any particular video uploader. In addition, we calculated the similarity of different videos by applying a vector space model and selected only one video from sets of overly similar videos.

Conclusions

In this study, we explored the use of existing resources and NLP technology to support listening item generation for the TOEIC Listening test. Good items need to use authentic situations and language in a wide variety of contexts. However, creating items for less familiar topics is a challenging task for item writers. As a result, the item writers tend to create items for familiar topics, and this can result in an imbalance in contexts and vocabulary. Most item writers tend to have expertise in the education and English language fields, which leads to overlap in experience from which to draw item ideas. To address this issue, we developed an automated seed video retrieval system, a list of vocabulary, and a word similarity tool. To examine the usefulness of these tools, we conducted a small-scale pilot study. Seven item writers created TOEIC Listening items using our tools and responded to a survey and interview on the last day of the pilot study. We evaluated the usefulness and impact of these tools on item generation based on the survey responses. In general, all tools were considered useful, and the word similarity tool in particular was rated the most useful. The preference of particular resources may vary across different item writers. The word similarity tool was most favored overall (four novice item writers), and the seed video collection was most useful to one of the experienced item writers. This finding suggests potential differences in the usefulness of tools between experienced item writers and new item writers. In our future exploration of this topic, we will extend our study and further investigate the impact of our resources with experienced item writers.

References

Capel, A. (2010). A1 – B2 vocabulary: Insights and issues arising from the English profile wordlists project. *English Profile Journal, 1*. https://doi.org/10.1017/S2041536210000048

Chen, M.-H., Huang, S.-T., Chang, J. S., & Liou, H.-C. (2015). Developing a corpus-based paraphrase tool to improve EFL learners' writing skills. *Computer Assisted Language Learning, 28*(1), 22 – 40. https://doi.org/10.1080/09588221.2013.783873

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(30), 213 – 238. https://doi.org/10.2307/3587951

Fuentes, A. C. (2002). Exploitation and assessment of a business English corpus through language learning tasks. *ICAME Journal*, *26,* 5 – 32.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers, 36*(2), 193 – 202. https://doi.org/10.3758/BF03195564

Heilman, M., & Madnani, N. (2012). A unified resource for distributional lexical similarity (Unpublished manuscript).

Heilman, M., & Smith, N. A. (2010). Good question! Statistical ranking for question generation. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 609 – 617). Stroudsburg, PA: Association for Computational Linguistics.

Heitler, D. (2005). *Teaching with Authentic Materials*. Retrieved from http://www.pearsonlongman.com/intelligent_business/images/teachers_resourse/pdf4.pdf

Hoshino, A., & Nakagawa, H. (2007). Sakumon: An assistance system for English cloze test. In R. Carlsen, K. McFerrin, J. Price, R. Weber, & D. A. Willis (Eds.), *Proceedings of Society for Information Technology & Teacher Education International Conference 2007*. San Antonio, TX: Association for the Advancement of Computing in Education.

Huang, Y.-T., Chen, M. C., & Sun, Y. S. (2012). Personalized automatic quiz generation based on proficiency level estimation. *Proceedings of the 20th International Conference On Computers In Education* (pp. 553 – 560). Retrieved from http://autoquizhttp.iis.sinica.edu.tw/docs/personalized.pdf

Huang, Y.-T., Tseng, Y.-M., Sun, Y. S., & Chen, M. C. (2014). TED quiz: Automatic quiz generation for TED talks video clips to assess listening comprehension. *Proceedings of the 14th IEEE International Conference on Advanced Learning Technologies* (pp. 350–354). Piscataway, NJ: IEEE. https://doi.org/10.1109/ICALT.2014.105

Liou, H.-C., Chang, J. S., Chen, H.-J., Lin, C.-C., Liaw, M.-L., Gao, Z.-M., & You, G.-N. (2013). Corpora processing and computational scaffolding for a web-based English learning environment: The CANDLE project. *CALICO Journal, 24*(1), 77–95.

Sheehan, K. M., Kostin, I., Napolitano, D., & Flor, M. (2014). The TextEvaluator tool. *The Elementary School Journal, 115*(2), 184–209. https://doi.org/10.1086/678294

Suggested citation:

Yoon, S.-Y., Lee, C. M., Houghton, P., Lopez, M., Sakano, J., Loukina, A., … Madnani, N. (2017). *Analyzing item generation with natural language processing tools for the* TOEIC® *Listening test* (Research Report No. RR-17-52). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12183

Action Editor: Beata Beigman Klebanov

Reviewers: Ikkyu Choi and Michael Flor