*ets research institute

MARCH
2026
RR–26-02

RESEARCH REPORT

Toward an Automatic Method for Generating Topical Vocabulary Test Forms for Specific Reading Passages

AUTHORS

Michael Flor, Zuowei Wang, Paul Deane, and Tenaha O’Reilly

ETS Research Report Series

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.

Toward an Automatic Method for Generating Topical Vocabulary Test Forms for Specific Reading Passages

Michael Flor, Zuowei Wang, Paul Deane, & Tenaha O’Reilly
ETS Research Institute, Princeton, New Jersey, United States

Abstract

Background knowledge is typically needed for successful comprehension of topical and domain-specific reading passages, such as in the STEM domains. However, there are few automated measures of student knowledge that can be readily deployed and scored in time to make predictions on whether a given student will likely be able to understand a specific content-area text. In this research report, we present our effort in developing the K-tool, an automated system for generating topical vocabulary tests that measure students’ background knowledge related to a specific text. The system automatically detects the topic of a given text and produces topical vocabulary items based on their relationship with the topic. This information is used to automatically generate background knowledge forms that contain words that are highly related to the topic and share similar features but do not share high associations to the topic. Prior research has indicated that performance on such tasks can help determine whether a student is likely to understand a particular text based on their knowledge state. The described system is intended for use with middle and high school student populations of native speakers of English. It is designed to handle single reading passages and is not dependent on any corpus or text collection. Here, we describe the system architecture and present an initial evaluation of the system outputs.

Keywords: topics, topical analysis, vocabulary, background knowledge, automatic item generation, AIG, assessment, reading comprehension, vocabulary recognition test

Corresponding author: Michael Flor, E-mail: mflor@ets.org

Acknowledgments

The lexicon of multi-word expressions is based in part on foundational research funded by the U.S. Department of Education Institute of Education Sciences (IES) though grant R305A080647. Previous research on background knowledge and reading comprehension was conducted with funding from the IES though grants R305A150176 and R305F100005. The opinions expressed are those of the authors and do not represent views of the IES. The specific ideas and designs for the fully automated tool described in this report were not part of any previous research.

Introduction

This research report describes the development of a prototype automated tool for the generation of vocabulary tests from topical reading passages. In Section 1, we present the context and motivation for developing this tool. Section 2 presents the tool and the computational linguistic aspects that power its capabilities. Section 3 presents an evaluation study that assessed the acceptability of generated test items. Discussion and limitations are presented in Section 4.

Background and Context

Why Measure Students' Background Knowledge in the Context of Reading?

The importance of background knowledge for successful reading comprehension has been emphasized by reading researchers for a long time (Castles et al., 2018; Cromley & Azevedo, 2007; Kintsch, 2004; Ozuru et al., 2009; Shapiro, 2004; Simonsmeier et al., 2021; Smith et al., 2021). In short, students with more relevant knowledge on a topic tend to have better comprehension than students with less relevant knowledge. Although background knowledge plays a key role in reading comprehension, it is rarely assessed when comprehension is measured. We argue that there is potential added value for measuring students' knowledge before they read a text, in either an assessment context or a non-assessment context. In an assessment context, a measure of background knowledge could help contextualize a reading score. A low score on a comprehension test may signal difficulties in understanding, or a low level of background knowledge may impede understanding. Having a measure of knowledge could help disentangle these two interpretations and improve reporting, as well as support the validity of a reading test.

In a non-assessment context like a regular classroom setting, a teacher may want to know which students may be at risk of not understanding a text that the class is about to cover. Such students could be given pre-reading activities that increase their knowledge, such as watching a video on the topic, reading a brief summary to provide them with background information, or a list of "must-know words" that are central to understanding the text. The aim of this research report is to describe the development of the prototype tool and describe the evaluation (item acceptance rate) for the generated test forms.

How to Measure Background Knowledge Efficiently?

There is a wide range of research on the role of background knowledge in reading comprehension. From a theoretical perspective, the Construction-Integration model is a good example of why knowledge is important for comprehension. In the Construction-Integration model (Kintsch, 2004), the reader's representation of a text, called the *situation model*, crucially depends on integrating background knowledge and the text contents. This is in part because texts are often incomplete and therefore require the reader to draw knowledge-based inferences to fill in gaps when forming a coherent situation model.

Despite the important role of background knowledge in comprehension, it can be challenging to measure efficiently. For instance, assessment of content and background knowledge at school has traditionally relied on propositional knowledge, assessed with factual statements, such as multiple-choice questions. Development of such items often requires expertise in the specific topic and is time consuming. For example, in the study reported by Cromley and Azevedo (2007), researchers exerted considerable effort to analyze the passages in a reading comprehension test and to identify the relevant background knowledge that would be important for students' understanding of the content. With advances in technology, natural language processing (NLP) systems can now successfully generate multiple-choice questions automatically from texts (Kurdi et al., 2020; Mulla & Gharpure, 2023). However, on-demand generation of assessments for measuring the relevant background knowledge for a specific passage remains an active research area.

Converging evidence from reading research has emphasized the central role of vocabulary knowledge in reading comprehension, for both native speakers (McKeown et al., 2017; Pearson et al., 2012; Perfetti & Stafura, 2014) and second language learners (Schmitt et al., 2011). In addition to general vocabulary, researchers have emphasized the importance of topical knowledge (Wang et al., 2021) and discipline-specific or topic-specific vocabulary for comprehension of reading materials at school (Fisher & Frey, 2014; Nagy & Townsend, 2012).

The intersection between background knowledge and vocabulary knowledge provides a good opportunity for efficiently assessing topical knowledge via a vocabulary test. For example, Stahl (2008) described a vocabulary recognition test (VRT) as an experimenter-constructed task for estimating vocabulary recognition in a content area (e.g., a text about insects). In that study, the test form consisted of 25 words; of these words, 18 were related to the content of the informational text, and 7 were used as distractors. The task was administered with paper and pencil to second-grade students in a school in the United States. The students had to circle only the words related to the topic, as a kind of yes/no task. As Stahl and Bravo (2010) have noted, such a test can be used as a pre-reading activity. It also can be used by a teacher to confirm that groups of students have similar levels of prior knowledge of a topic. However, students who are not familiar with the topical vocabulary are at risk of not comprehending the reading materials. A VRT-type of assessment has the potential to enable teachers to identify such students and provide them with vocabulary support before the reading activity.

Topical Vocabulary as an Efficient Proxy for Background Knowledge

There is evidence to suggest that a topical vocabulary test can be used as an efficient estimate of the level of students' background knowledge. For instance, O'Reilly et al. (2019) investigated the use of a similar yes/no test format for students in Grades 9–12, with reading materials about ecosystems. Students were presented with a list of terms and were asked to indicate whether each word was related or unrelated to the topic of ecology. The test form included 44 words, of which 26 were topical words and 18 were distractors. Only nine of the topical words were explicitly mentioned in the text. Other topical words were manually procured by a test developer, but this was facilitated with suggestions derived from an NLP-generated database of word associations.

O'Reilly et al. (2019) demonstrated that such a topical vocabulary test was a viable measure of background knowledge and could predict students' reading comprehension success. It was also efficient because the knowledge test took only 5–7 minutes to complete. Moreover, the test helped identify students who lacked topical knowledge to achieve adequate comprehension. Specifically, the experimenters were able to numerically identify a knowledge threshold such that students who scored below the knowledge threshold were not likely to comprehend a subsequent text related to the topic of the knowledge test, as compared to students who scored above the knowledge threshold. In addition, they also found that knowledge (or lack of thereof) of only the six most strongly related topical words provided an indication of whether a student would be above or below the knowledge threshold. The group below the threshold had an average accuracy of 64% on those words, while the group above the threshold had an average accuracy of 95%. This suggests that a topical vocabulary test can aid teachers in identifying students who may struggle when reading a given passage because of their overall level and unfamiliarity with must-know words related to the topic of the text. Interestingly, these must-know words included not only words from the passage but also words that *did not* appear in the passage, supporting the notion that the network of associations is important, as is the vocabulary itself. Notably, not only is the test useful for identification but it may have instructional implications (e.g., in some cases, teachers may choose to teach the must-know words before reading, to promote better comprehension outcomes).

Given the advantages of a topical vocabulary recognition test, it may serve as a practical and efficient formative assessment tool to be used by teachers. A crucial step in that direction is to conceptualize a tool that can generate such test forms automatically. In this report, we present a first iteration of the K-tool, an automatic item generation (AIG) system that was designed to generate such tests for domain-specific reading passages. Section 2 presents the system design and implementation details. Section 3 describes an initial evaluation of the system, focusing on the acceptability of the generated test items.

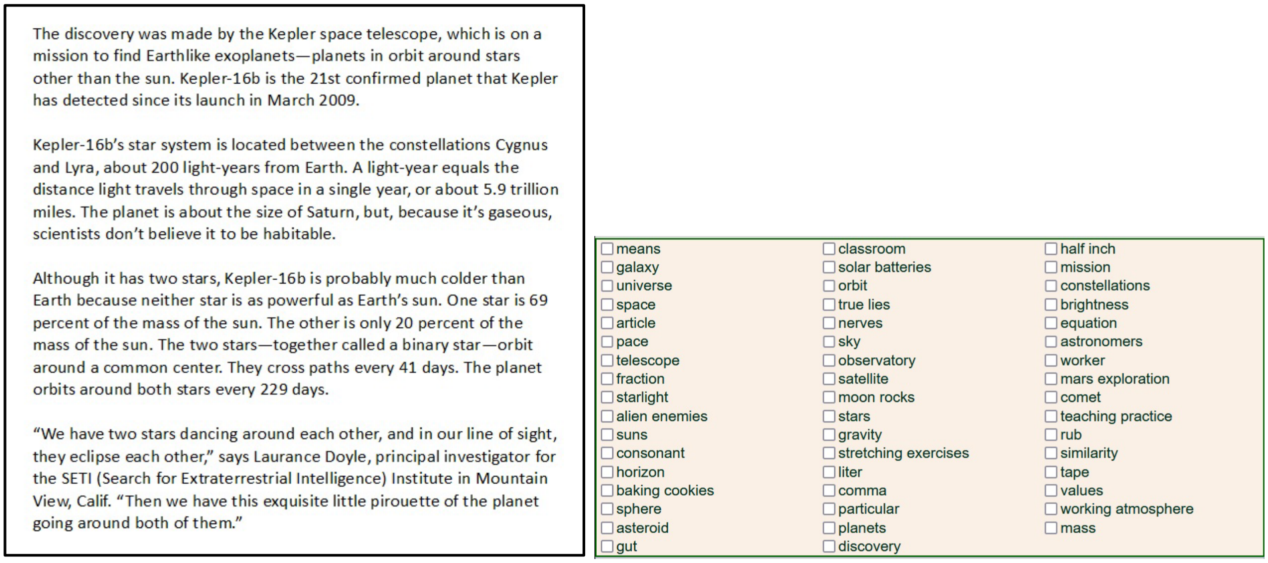
The K-tool System

Test Form Composition

The K-tool system was designed to take a single text passage as input and generate one or more topical vocabulary test forms for it. In this study, we focused on passages between 400 and 1,500 words, lengths typical for high school students to read in one sitting. Our prior research (O’Reilly et al., 2019; Wang et al., 2021) has demonstrated that a topical vocabulary test with approximately 30–50 items would achieve a reliability of .90 (Cronbach’s alpha). As such, the current K-tool system aims to generate 50 items per passage: 14 terms would be in-document topical terms (TID); 14 would be out-of-document topical terms (TOD), that is, from an external lexicon; and 22 would be non-topical (NT) terms, which would serve as distractors (also obtained from an external lexicon). The quantity of 50 and the 14:14:22 composition were set as initial targets for developing the current system. In the future, such settings might be left for users to choose, and studies can be conducted on finding optimal settings (see the Further Extensions subsection in the Discussion section). Instructions on a topical vocabulary test form would ask the test taker to indicate whether each term (e.g., *galaxy*, *tape*) is related to the given topic (e.g., astronomy). The proportion of correctly marked terms (accepted topical words and rejected non-topical words) would indicate a student's familiarity with the vocabulary of the given topic. A sample text and a corresponding vocabulary test form are shown in Figure 1. Note that the reading passage itself is *not* part of the test form, because the test is intended to assess topical knowledge *prior* to the reading activities.

At present, we require that all test terms on a form be nouns or noun phrases. We focus on nouns because nouns typically constitute the most prominent topical vocabulary for texts. We leave utilization of verbs and adjectives in testing for future research—this aspect will be informed by further exploring educational needs. However, this does not mean that we do not extract topical verbs and adjectives from a text; we just do not include them on a generated test form. The system also uses multi-word expressions (MWEs), because nominal MWEs often represent important topical terminology, such as *potential energy* or *greenhouse gas*.

Figure 1. A Text Passage (Part Shown) and a Corresponding Topical Vocabulary Test Form



The discovery was made by the Kepler space telescope, which is on a mission to find Earthlike exoplanets—planets in orbit around stars other than the sun. Kepler-16b is the 21st confirmed planet that Kepler has detected since its launch in March 2009.

Kepler-16b's star system is located between the constellations Cygnus and Lyra, about 200 light-years from Earth. A light-year equals the distance light travels through space in a single year, or about 5.9 trillion miles. The planet is about the size of Saturn, but, because it's gaseous, scientists don't believe it to be habitable.

Although it has two stars, Kepler-16b is probably much colder than Earth because neither star is as powerful as Earth's sun. One star is 69 percent of the mass of the sun. The other is only 20 percent of the mass of the sun. The two stars—together called a binary star—orbit around a common center. They cross paths every 41 days. The planet orbits around both stars every 229 days.

"We have two stars dancing around each other, and in our line of sight, they eclipse each other," says Laurance Doyle, principal investigator for the SETI (Search for Extraterrestrial Intelligence) Institute in Mountain View, Calif. "Then we have this exquisite little pirouette of the planet going around both of them."

□ means
□ galaxy
□ universe
□ space
□ article
□ pace
□ telescope
□ fraction
□ starlight
□ alien enemies
□ suns
□ consonant
□ horizon
□ baking cookies
□ sphere
□ asteroid
□ gut
□ classroom
□ solar batteries
□ orbit
□ true lies
□ nerves
□ sky
□ observatory
□ satellite
□ moon rocks
□ stars
□ gravity
□ stretching exercises
□ liter
□ comma
□ particular
□ planets
□ discovery
□ half inch
□ mission
□ constellations
□ brightness
□ equation
□ astronomers
□ worker
□ mars exploration
□ comet
□ teaching practice
□ rub
□ similarity
□ tape
□ values
□ working atmosphere
□ mass

Note. The instruction for this test form would be "Select all of the terms that are related to the topic of astronomy."

High-Level Process for Creating Forms

With those requirements, the conceptual design of an AIG system is as follows:

Step 1. Identify the major topic of the text.

Step 2. Select all words (and nominal MWEs) in the text that are strongly related to the main topic.

Step 3. Select nouns (and nominal MWEs) from a general English vocabulary that do not appear in the text but are strongly related to the major topic.[1]

Step 4. Select nouns (and nominal MWEs) from a general English vocabulary that do not appear in the text and that are absolutely not related to the major topic(s) of the text. This set would be used to generate distractors.

Finding topics in texts is a venerable research area in NLP. It has long been dominated by Latent Dirichlet Allocation (LDA) and related approaches (Vayansky & Kumar, 2020) and recently by neural topic modeling approaches (Wu et al., 2024). However, topic modeling has focused on automatically identifying topics in large collections of texts, and those methods are not directly applicable to analysis of a stand-alone text document. We are interested in identifying the topics in any stand-alone document, without relying on prespecified taxonomies of topics or on any collections of documents.

Another relevant and venerable area of NLP research is automated extraction of keywords and keyphrases (Xie et al., 2023). Such approaches can operate on a single text. However, they are often limited to extracting very few representative keywords. Chau et al. (2021) described an approach to extracting key concepts from a large textbook, utilizing keyphrase extraction methods, but also relying on the structure of chapters in a textbook. However, keyword extraction is not a suitable approach in our case, because keywords may reflect different subtopics in a text, and keywords do not exhaustively reflect all the topical words in a text. We need to find in a text all the words (types, not tokens) that represent the main topic of the text.

We introduce a new method for topic detection in a single text passage. From topic modeling literature, we borrow the convenient definition of a *topic* as a set of strongly related words. The outline of our approach is as follows: (a) cluster the words of the text by semantic similarity and (b) find the cluster (or set of clusters) that is most representative of the text. The words in those clusters would be the vocabulary of the main topic of the text. To implement these steps, we utilize neural embeddings for both words/phrases and whole texts.

Preprocessing Steps

The text of the reading passage is preprocessed as follows: part of speech (POS) tagging, MWE detection, vectorization of the whole text, and also vectorization of each word or MWE term.

POS Tagging

We identify the POS for every word in the text. This is used for identifying the nouns, verbs, adjectives, and other major POS types in a text. We utilize the OpenNLP POS tagger.[2]

MWE Detection

Given the text passage, we perform noun phrase MWE detection using an in-house predeveloped lexicon of 68,000 nominal MWEs (e.g., *space shuttle*, *abdominal fat*).[3]

Whole-Text Vectorization

The reading passage text is embedded into a single vector representation. This vector serves as a representation of the semantic content of the whole text and is used later for estimating semantic relatedness of various components to the whole document. We use the MiniLM-L6-v2 model from the Sentence-BERT library (Reimers & Gurevych, 2019). Since BERT models (and Sentence-BERT models) have limits on the number of word tokens (text length) that they can encode (e.g., 400 words), we break longer texts into chunks, vectorize every chunk and then combine the resulting vectors by computing the average vector—a single vector for the whole document (Iso et al., 2021; Sannigrahi et al., 2023; Sun & Nenkova, 2019). The document vector is L2 normalized.[4]

Term Vectorization

We use a precomputed dictionary of static embeddings for a lexicon of 150,000 English words. The dictionary was developed by using Sentence-BERT as an embedder, with the same MiniLM-L6-v2 model. In the same way, we also developed static embedding vectors for our list of 68,000 MWEs. Using those resources, every word or MWE for a given text is vectorized via simple dictionary lookup. All vectors are normalized with L2 normalization. The need for static acontextual embeddings is motivated, as we also use the same vectors to check how strongly words from an *external list* are related to the topic of a given document (see below). Note that the vector for the document and the vectors for all terms (in the document and external) must be from the same vector space model.

Finding Topics in the Text

To find topical groups of terms in the text, we perform clustering of the words and MWEs of the document using cosine similarity between embedding vectors as the clustering similarity measure.

Because we are interested in topical analysis of terms, we utilize nouns, verbs, and adjectives and filter out certain types of words that are not useful for such analysis. We use POS tags and stop lists to filter out determiners, prepositions, function words, numbers expressed in digits, and also adverbs, proper nouns (names), interrogatives, demonstratives, do/be/have verb forms, and modal verbs. We include detected MWEs as units but exclude their parts. For example, if *space shuttle* is included, its constituent *shuttle* is excluded, unless this word appears by itself elsewhere in the text. Only word types are used for clustering and a word form that appears multiple times in text would be included only once. However, we do keep track of how many times each word form appears in the text.

Since the number of topical term groups in any given text is not known a priori and can vary from text to text, we opt to use a clustering approach that does not require prespecifying the number of clusters. The affinity propagation clustering (Frey & Dueck, 2007) works well for our purposes, as it automatically finds the optimal number of clusters for each text. For words (and MWEs) of a given text, we run affinity propagation clustering using vector cosine similarity as the similarity measure until convergence for 10 iterations and collect the resulting term clusters.

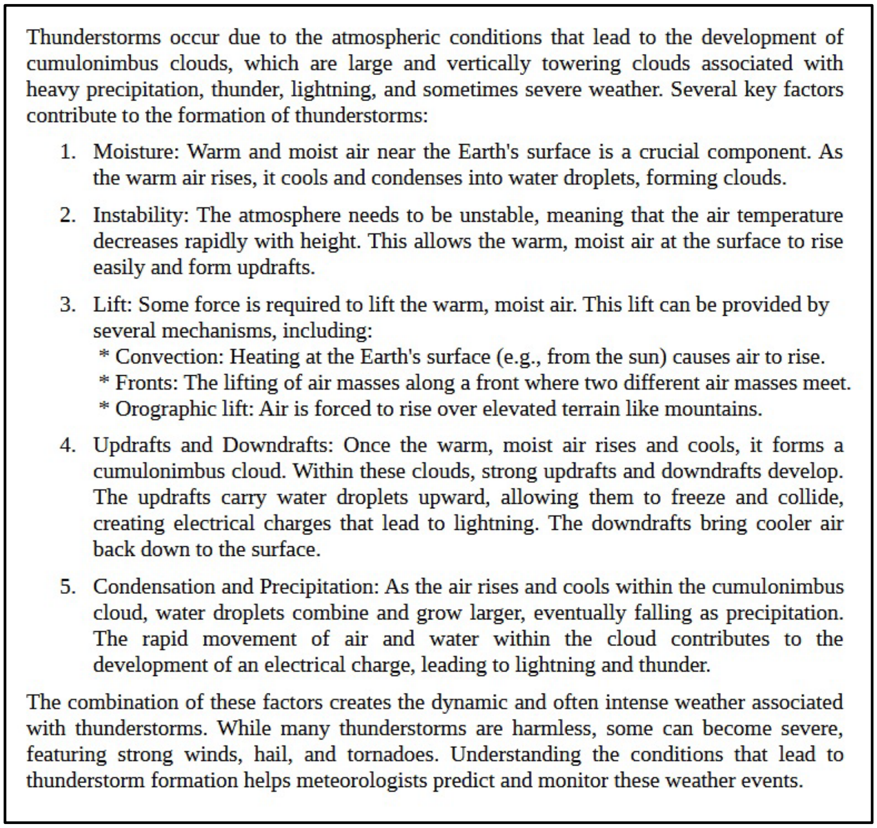
For each cluster, we compute a centroid vector that will represent that cluster. The centroid is computed as a weighted average of cluster term vectors, where the weights are the counts of the terms in the text (aka *tf* weighting). The centroid vector is also L2 normalized.

After those steps, we can estimate which term clusters represent the more central topics of the text. We compute cosine similarity between the vector of the whole document and the centroid vector of each cluster, and then sort the clusters by their similarity to the whole document. Clusters that are most similar to the whole document represent the central, most important topics of the document. To illustrate this point, we present clusters that were generated for a 314-word-long informational passage about thunderstorms (Table 1; Figure 2).

Table 1. The Top Seven Clusters for Terms From a Text About Thunderstorms, Sorted by Cluster Cosine Similarity to the Whole Text

Cluster number	Cluster's cosine to document	Terms
1	.531	thunder, thunderstorms, severe weather, clouds, hail, winds, tornadoes, lightning, meteorologists
2	.347	condenses, precipitation, water, droplets, moisture, moist air
3	.325	atmosphere, conditions, atmospheric conditions, weather, airmasses
4	.242	featuring, unstable, collide, meet, events, causes, occur, cumulonimbus
5	.192	warm air, air, sun, heating, air temperature, warm, freeze, cooler air
6	.163	elevated, carry, lift, lifting, rises, movement, force
7	.162	surface, terrain, mountains

Figure 2. (a) A Text Passage About Thunderstorms and Two Generated Topical Vocabulary Test Forms: (b) Easy and (c) Difficult

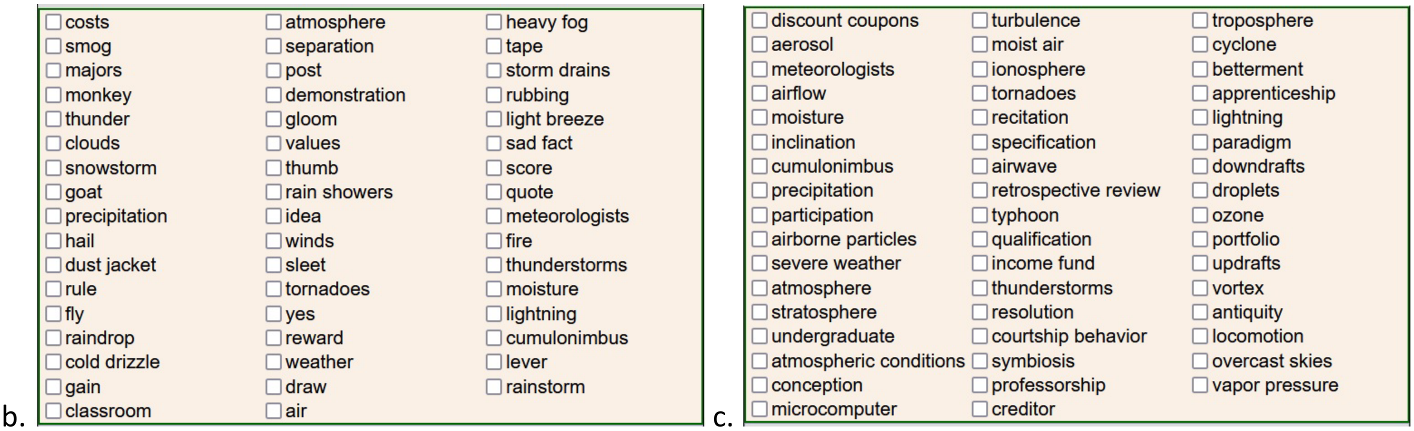


Thunderstorms occur due to the atmospheric conditions that lead to the development of cumulonimbus clouds, which are large and vertically towering clouds associated with heavy precipitation, thunder, lightning, and sometimes severe weather. Several key factors contribute to the formation of thunderstorms:

1. Moisture: Warm and moist air near the Earth's surface is a crucial component. As the warm air rises, it cools and condenses into water droplets, forming clouds.
2. Instability: The atmosphere needs to be unstable, meaning that the air temperature decreases rapidly with height. This allows the warm, moist air at the surface to rise easily and form updrafts.
3. Lift: Some force is required to lift the warm, moist air. This lift can be provided by several mechanisms, including:
   * Convection: Heating at the Earth's surface (e.g., from the sun) causes air to rise.
   * Fronts: The lifting of air masses along a front where two different air masses meet.
   * Orographic lift: Air is forced to rise over elevated terrain like mountains.
4. Updrafts and Downdrafts: Once the warm, moist air rises and cools, it forms a cumulonimbus cloud. Within these clouds, strong updrafts and downdrafts develop. The updrafts carry water droplets upward, allowing them to freeze and collide, creating electrical charges that lead to lightning. The downdrafts bring cooler air back down to the surface.
5. Condensation and Precipitation: As the air rises and cools within the cumulonimbus cloud, water droplets combine and grow larger, eventually falling as precipitation. The rapid movement of air and water within the cloud contributes to the development of an electrical charge, leading to lightning and thunder.

The combination of these factors creates the dynamic and often intense weather associated with thunderstorms. While many thunderstorms are harmless, some can become severe, featuring strong winds, hail, and tornadoes. Understanding the conditions that lead to thunderstorm formation helps meteorologists predict and monitor these weather events.

a.



b.

□ costs
□ smog
□ majors
□ monkey
□ thunder
□ clouds
□ snowstorm
□ goat
□ precipitation
□ hail
□ dust jacket
□ rule
□ fly
□ raindrop
□ cold drizzle
□ gain
□ classroom
□ atmosphere
□ separation
□ post
□ demonstration
□ gloom
□ values
□ thumb
□ rain showers
□ idea
□ winds
□ sleet
□ tornadoes
□ yes
□ reward
□ weather
□ draw
□ air
□ heavy fog
□ tape
□ storm drains
□ rubbing
□ light breeze
□ sad fact
□ score
□ quote
□ meteorologists
□ fire
□ thunderstorms
□ moisture
□ lightning
□ cumulonimbus
□ lever
□ rainstorm

c.

□ discount coupons
□ aerosol
□ meteorologists
□ airflow
□ moisture
□ inclination
□ cumulonimbus
□ precipitation
□ participation
□ airborne particles
□ severe weather
□ atmosphere
□ stratosphere
□ undergraduate
□ atmospheric conditions
□ conception
□ microcomputer
□ turbulence
□ moist air
□ ionosphere
□ tornadoes
□ recitation
□ specification
□ airwave
□ retrospective review
□ typhoon
□ qualification
□ income fund
□ thunderstorms
□ resolution
□ courtship behavior
□ symbiosis
□ professorship
□ creditor
□ troposphere
□ cyclone
□ betterment
□ apprenticeship
□ lightning
□ paradigm
□ downdrafts
□ droplets
□ ozone
□ portfolio
□ updrafts
□ vortex
□ antiquity
□ locomotion
□ overcast skies
□ vapor pressure

Note. The instruction for these test forms could be "Select all of the terms that are related to the topic of weather conditions."

Creating Sufficient Vocabulary for Multiple Test Forms

Our current design for a single topical vocabulary test form requires 50 terms (all of them nouns or nominal MWEs), of which 14 topical terms are from the text, 14 topical terms are from the external lexicon, and 22 topically unrelated terms are used as distractors. These requirements are set for a *single* test form. However, in some cases, it might be beneficial to generate more than one test form. For example, we may want to have both an easy and a more difficult form to accommodate larger variations in the level of student knowledge (the rationale is presented in the next subsection, Generating Multiple Test Forms). To allow for the generation of multiple test forms, we need to oversample the vocabulary.

With this in mind, to begin generating a test form for a text, we need to select the required number of topical terms from the top clusters that were obtained. To allow for more flexibility, we introduce a more general term-selection process that allows for generating multiple test forms for a given text. The idea is to overselect terms into a pool of accepted terms from which multiple test forms can be generated. The process has three steps: selecting (a) topical terms from the text, (b) topical terms from the external general vocabulary, and (c) the distractor terms.

We generate a pool of topical terms from the text by overselecting nominal terms from the top topical clusters (usually the three top clusters). By default, we select 25 nominal terms from those clusters. The number obviously needs to be higher than the 14 TID terms for a form. A good quota might be 28, but for some texts, 28 TID terms are not always available; thus, we set the initial quota to 25. This setting is likely to be modified in the future (see Further Extensions subsection in the Discussion section).

The next step is to select topical terms from outside of the given text, and to select non-topical terms to be used as distractors. For this we use a general list of English nouns (~30,000 lemmas, a subset of the 150,000 general list) and the list of MWEs (~68,000 entries) as the supply lists from which candidates are drawn. For both lists, we have corresponding embedding vectors prearranged as general precomputed resources that are used by the K-tool.

Selecting terms for an out-of-document topical list amounts to (a) scanning the vocabulary supply lists, (b) filtering out any terms that already appear in the text, and (c) selecting words that have sufficient semantic similarity to the in-document topical terms. For example, if a document has the words *car*, *road*, and *drive* among its top topical terms, we may wish to bring from the external list such words as *driver*, *passenger*, *motor*, and *traffic*.

The notion of "sufficient semantic similarity" requires some elaboration. One approach could be to find a certain threshold value of similarity to be used as a cutoff. However, we opted for a more dynamic approach. We sort the out-of-document terms by their similarity to the major topics of the document and pick the *n* top-rated terms as needed. For this we define a pool of top topical-in-document terms, which is simply the list of all terms from the three top term-clusters of the document. An external candidate term needs to be semantically similar or semantically related to this pool of terms.

The measure for such computation could be simply the cosine similarity between vectors of candidate terms (external list) and the vectors of the in-document topical terms, or through word cooccurrences. Both approaches are well known in the computational linguistic literature. In this study, we combined the cosine similarity measure with the cooccurrence measure. A first-order cooccurrence-based approach reflects the notion that words that frequently occur together are topically related (Schütze & Pedersen, 1997). The second-order distributional similarity approach reflects the notion that words occurring within similar contexts are semantically similar or related (Bullinaria & Levy, 2012; Lin, 1998; Turney & Pantel, 2010). Although some studies have compared the two approaches (Liebeskind et al., 2018; Purandare & Pedersen, 2004), their combination has also been explored (Flor et al., 2019). Vector cosine helps emphasize semantic similarity (*car* and *truck*), while cooccurrence helps emphasize semantic relatedness (*car* and *wheel*).

In the current work, for each candidate vocabulary term, we compute its support in the document as follows:

$$Support(T_c) = \sum_{j=0}^{n} \left(\left(cosine(T_c, T_{dj}) + PNPMI(T_c, T_{dj}) \right) \times \, log_{10}\left(count(T_{dj}) + 1 \right) \right)$$

where T_c is a new candidate term, T_{dj} is the jth word from the list of N top in-document topical terms, and $count(T_{dj})$ is the number of occurrences of T_{dj} in the document.[5] PNPMI is positive normalized pointwise mutual information, which is a variant of the well-known pointwise mutual information (PMI) measure. PMI is defined as follows (Church & Hanks, 1990):

$$\text{PMI} = log_2 \frac{p(a,b)}{p(a) \times p(b)} \ ,$$

where a and b are words, $p(a)$ and $p(b)$ are probabilities of each word, and $p(a,b)$ is the probability of joint occurrence. The probabilities are estimated from counts in a very large corpus. We used a language model of word cooccurrence within paragraphs, trained on a corpus of more than 2 billion words (Flor & Beigman Klebanov, 2014). Normalized PMI (NPMI) (Bouma, 2009) has values constrained in the range (-1, 1) and is defined as

$$\text{NPMI} = \frac{log_2 \frac{p(a,b)}{p(a) \times p(b)}}{- \, log_2\left(p(a,b)\right)} \ .$$

PNPMI takes the value of NPMI, or zero if NPMI is negative or if the value for the cooccurrence of words a and b is not available in the database. When the term a or b (or both) is an MWE, we compute the association as an average PNPMI value between the words of the first term and the words of the second term.

Generating Multiple Test Forms

As stated earlier, our design allows for generating more than one test form per text. At this stage in system development, we decided to differentiate the generated test forms by their collective difficulty of vocabulary terms. During the term-selection process, for each term, we retrieve its grade-level estimator (a decimal value) as a measure of its difficulty. Thus, each term in our pool of selected topical terms also has a grade-level value. Then, during form assembly, we select from this pool the terms with the lowest grade-level values, thus generating a relatively “easy” form, or the terms with the highest grade-level values, thus generating a relatively “difficult” form. Grade-level estimators for single words come from the VXGL resource, a list of 126,000 English words mapped to their estimated grade levels (Flor et al., 2024). For multi-word expressions, we do not yet have a predefined resource; therefore, we estimate the grade level as follows: find the grade level for each constituent word, pick the word with the highest value, increase that value by 25%, and use the result as the grade-level estimate for the phrase.[6]

Another consideration for form assembly is that the distribution of grade levels among the topical terms selected from outside-of-the-document (TOD), and also among the distractor non-topical terms (NT), should be approximately equivalent as for the topical-in-document terms (TID). Otherwise, many of the terms on a form may stand out as being too easy or too difficult relative to the other terms. For example, we want to avoid a case in which most of the distractors are “too easy” (very familiar words). To alleviate this, we use the distribution of grade levels for the TOD terms as our guiding distribution. As noted, when selecting terms from the external lexicon, we can create rather large pools of suitable candidates (more than only 14 and 22, respectively). We then perform a second round of selection from those large pools, ensuring that the distribution of grade levels for terms in those pools approximately matches the distribution of grade levels in the TID pool. Then, during generation of a test form, we can pick the terms with the highest (or lowest) grade-level values from each of the TID, TOD, and NT pools, knowing that the grade-level distributions in them are already matched. A sample text and two corresponding autogenerated vocabulary test forms are shown in Figure 2.

The rationale for generation of different test forms is related to the overall estimation of passage and test difficulty. Passage text complexity can be estimated with a variety of methods, such as readability formulas (e.g., Flesch–Kincaid) or by using a specialized text complexity tool, such as the TextEvaluator® (Sheehan et al., 2014). However, the difficulty of a test form that is generated for a given passage is not the same as the overall difficulty of the passage. A vocabulary test form is just a collection of nouns (and nominal MWEs), and many of them are not even from the passage. Also, the core function of the K-tool item generator is to ensure *topical relations* of the included terms. Without additional control measures, the terms, especially those brought from the general lexicon, can vary in their complexity and grade-level suitability. For instance, the non-topical words could be some common, everyday words, or they could be quite uncommon words (but still non-topical). Also, given a topic, the topically related words can have considerable variability with respect to their difficulty or familiarity. For example, consider the topic “cars” with topical words *engine* (easy) versus *crankshaft* (hard). Thus, it might be a useful feature to indicate the difficulty of the vocabulary items (as a supplement to the test form). In addition, it might be desirable to have some control over the difficulty of the overall test form (e.g., relative to the original reading passage). Such flexibility might also be useful when the same passage is used for different grade levels.

Difficulty of the test form can be manipulated, in part, by controlling the estimated difficulty of the terms included on the form (while still keeping the topicality distinctions as needed). Vocabulary difficulty can be estimated by word frequency or, in our case, by the VXGL resource. Since we can control the estimated difficulty for the form, the next question is then what difficulty we want to impose. A natural anchor can be the range of difficulties for the topical words already included in the passage. This is a data-driven approach: It depends only on the text. Given such a range, we can then add topical and non-topical terms from the lexicon, with estimated grade levels closer to the easy/hard points of the range. Thus we can vary the estimated difficulty of the test form relative to the range of grade levels of the topical words in the text. The test form can be made relatively easy or relatively difficult. We believe that the ability to control the estimated difficulty of a vocabulary test form provides a flexibility feature for the prototype tool. Whether this feature will be considered useful by users of the K-tool is an open question for future research. It is also an open question whether the predictive accuracy of the test may vary by the predicted difficulty of the form. For example, a very easy form might be less predictive of student comprehension. This aspect would be investigated in future research.

Evaluation

This section describes an initial evaluation of the K-tool system. The purpose of this evaluation was to assess the acceptability of items that are generated by the system. This is known as *intrinsic evaluation*; it is concerned only with the adequacy of the outputs and not with the system’s use or efficacy in the classroom. An intrinsic evaluation is necessary to confirm that we have a viable prototype system prior to any potential future field tests that might be conducted with teachers and schools.

Method

Twenty reading passages were selected from ReadWorks.org, a nonprofit organization that provides school reading materials for Grades K–12 in the United States. All 20 passages were expository texts oriented on science, technology, engineering, and mathematics (STEM) topics for Grades 9–10 (physics, life sciences, technology) and ranged in length from 485 to 1,465 words (average 1,058 words). We focused on STEM because such passages are typically topical and usually have specifically topical vocabulary, and also because STEM passages usually introduce topics that are not widely familiar to K–12 students—the kinds of passages for which assessment of prior knowledge may be very relevant. We chose high school-level texts because this level is the most complex in the K–12 range.

For each passage, two vocabulary test forms were generated, of easy and hard relative difficulty, respectively. All vocabulary terms are nouns or noun phrases.

It turns out that for our texts, it was not always possible to produce 14 different TID terms (nouns) for the easy test form and 14 other TID terms for the difficult test form. As it can happen, the number of distinct topical nouns in a text, especially a shorter text, can be fewer than 28. In such cases, the system can still produce two forms per text, but the forms will share some of the topical terms. However, we also found that for 4 out of 20 texts, the K-tool system could not find even 14 in-document topical nouns; for those texts, it found 9, 10, 11, and 13 terms, respectively.

Owing to the scarcity of TID terms, we have the following complication. We would expect to obtain 2,000 terms in total: 50 terms per form, with 2 forms per document and 20 documents. However, we have only 1,971 total terms for evaluation. Out of those, 140 terms were shared by easy and difficult test forms, so we have 1,831 unique term-document cases for evaluation. (See more on scarcity in the Discussion section under the Technical Limitations subsection.)

Our evaluation was concerned with the acceptability of each term on its test form. It was carried out by two evaluators. One evaluator was an intern, a college senior studying psychology; the second evaluator was one of the authors, an experienced linguist. The protocol of annotation was as follows: Given a reading passage and a test form automatically generated for the text, the human evaluator’s task was to annotate each term on the form. The annotation for each term was a binary decision: yes/no acceptable. All three types of terms, topical-in-document, topical-out-of-document, and non-topical, were clearly identified as such (labels were automatically provided by the system). The criteria for acceptability were as follows: Topical terms had to be clearly related to the topic of the passage, whereas non-topical terms had to be clearly unrelated to the topic. The text passage was available to the raters during annotation.

The annotation task was defined by two of the authors. Prior to the main task, the raters conducted a training session (with two passages), and then a pilot annotation (on one passage), and achieved agreement of 94%.

Results

Of the 1,831 unique terms, evaluators agreed in 1,766 cases; that is, agreement of 96.45%. Cohen's kappa is .7495 (substantial agreement, according to Landis & Koch, 1977). For individual texts, agreement was relatively high, ranging from 92% to 99%. For the easy forms, agreement ranged between 90% and 100% across the 20 texts; for the difficult forms, agreement ranged between 89% and 100%.

Next, we consider term acceptability. We used a strict criterion of acceptance: A term was accepted only if *both* evaluators accepted it. With such a strict criterion, the term acceptance rate was 1,658 of 1,831, that is, 90.55%. Thus, a large majority of the automatically generated test terms were considered adequate for vocabulary testing purposes.

Table 2 breaks down the accepted terms by type and source. The acceptance of non-topical terms was high, 97%. The algorithm seems capable of supplying non-topical distractors for the vocabulary test. The system's ability to extract topical vocabulary from texts was even higher, 99.5%. However, the ability to supply topical terms from an external lexicon was less accurate, with an acceptance rate of 74.5%. A breakdown for each passage is given in the appendix.

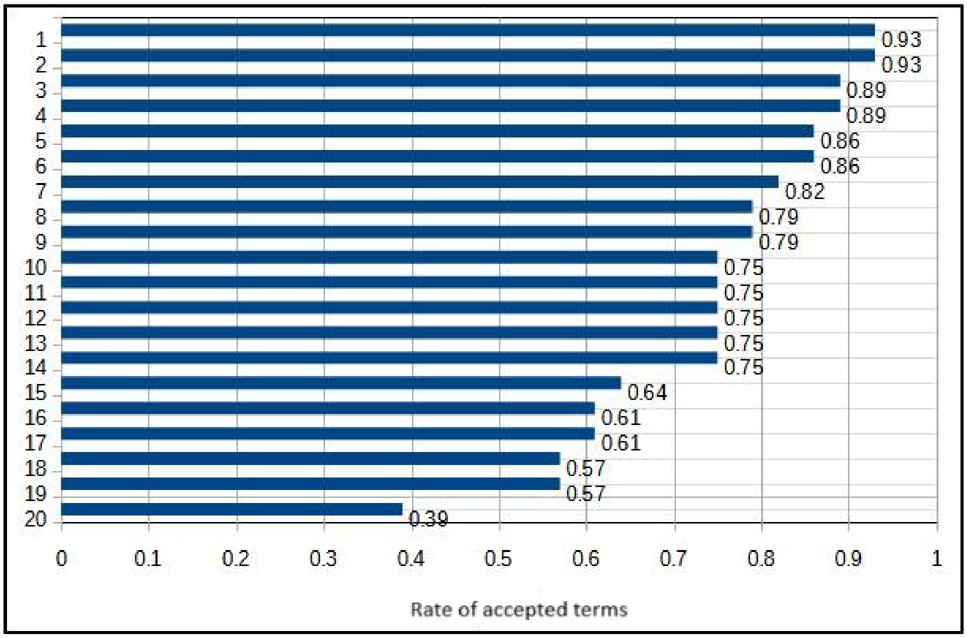
Table 2. Acceptance Rates for Evaluated Terms

Term type	Accepted	Total	Acceptance rate
NT: Non-topical (distractors)	852	880	.970
TID: Topical in-document	389	391	.995
TOD: Topical from lexicon	417	560	.745
Total	1,658	1,831	.906

Note. NT = non-topical; TID = topical terms from inside the document; TOD = topical terms retrieved from outside the document.

We set out to investigate the system performance for topical-out-of-document (TOD) terms. Figure 3 presents the acceptance rates for TOD terms in the test forms generated for 20 texts. For each text, we have 28 TOD terms, so the percentages are relative to that maximum number. Seven texts have acceptance rates greater than .8. One text in particular stands out with a very low acceptance rate (.39). That text is about scientists investigating sediment rocks that provide evidence about oxygen in the atmosphere on Earth millions of years ago. The topical terms from the document (TID) are a mix of atmosphere-related terms and chemistry-related terms, such as *air*, *waters*, *iron*, *powder*, *weathers*, *oxygen*, *mineral*, *crust*, *atmosphere*, *reactions*, *layers*, *compounds*, *sediments*, and *ozone*. The TOD terms include terms that are related to the topic via the chemistry theme but not directly relevant to the discussion in the text. Among the TOD terms rejected by evaluators are *fuel combustion*, *carbon monoxide poisoning*, *photosynthesis*, *gasoline*, *aerosol*, and *ammonia*. What we encounter in this case is that the topic of the text is at an *intersection of two domains* (atmosphere and chemistry) and that TOD words that are strongly related to only *one* of the domains might be too far from the relevant thematic intersection. Another text in our collection discusses biological aspects of the connection between dinosaurs and modern birds. The strict-criterion acceptance rate for TOD terms for that text was .64. Among the evaluator-rejected TOD terms for that text, we find *gastropod*, *cephalopod*, *mammal*, *mammoth*, and *aquatic plants*. It is easy to see the general biological relation, but evaluators considered those terms to be too far from the specific topic of the text.

Figure 3. Acceptance Rates for Topical-Out-of-Document Terms, for 20 Texts



Such analysis suggests that we may need to reconsider some of our initial assumptions. The topical vocabulary test is intended to test students' knowledge from two perspectives: vocabulary in a *given content area* and vocabulary related to a *specific reading passage*. But what is the relevant content area? If we define the content area too broadly (e.g., evolutionary biology), such a definition might be too general for a specific text. On the other hand, if we define the relevant content area too narrowly, we might be in a tight corner and may have difficulty finding enough relevant terms for a test. A discussion between the annotators revealed that they sometimes tended to focus on the relevance of a candidate TOD term to the actual text, rather than to a broader content area.

Moreover, we also need to consider cases in which texts intersect domains that are not usually discussed together. For example, one of the texts in our collection discussed sound phenomena that are encountered in a baseball game. The terminology was a mix of baseball-related terms and terms related to the physics of sound waves. The acceptance rate of TOD terms in this case was .57. This example illustrates the need to consider not only the scope and grain size of a content area (e.g., sound waves, ear anatomy, sound location) but also what is the main content area itself—is it baseball, or is it physics? If one sticks to the obvious—assuming it is both baseball and physics—then many TOD terms are rejected because they are associated with either baseball or physics, but not both. Such examples illustrate that we need to investigate further what kinds of content areas or domain knowledge we can measure with vocabulary tests. The issue is not vocabulary but rather what the relevant content area should be and how we can demarcate it for assessment purposes.

Discussion

Students' reading comprehension difficulties can originate from a variety of different sources. For example, our prior work has demonstrated that comprehension problems could be associated with students' poor decoding skills (Wang et al., 2024; Wang et al., 2020; Wang et al., 2019), being unfamiliar with the reading topic (O'Reilly et al., 2019; Wang et al., 2021), or inadequate reading fluency (Sabatini et al., 2019). Identifying the bottleneck of comprehension can inform instructional activities. Relevant to the current study, some studies have demonstrated that students' topical vocabulary is a good indicator of potential comprehension difficulties. O'Reilly et al. (2019) showed that student performance on a 5-minute topical vocabulary test, administered before students read materials on the topic, was a good indicator for whether students could achieve adequate comprehension when reading. Furthermore, students' topical knowledge can serve as an indicator of the cultural relevance of the reading materials (Wang et al., 2025).

We envision that a topical vocabulary test can be an important component in reading instruction. A low score on a traditional text-comprehension test usually indicates a problem, but it does not indicate the source of the problem. Measuring background knowledge allows one to identify one of the major potential causes of low comprehension. In such a context, the K-tool is aimed at generating topical vocabulary tests that could be administered before the actual reading. Such tests can be quick to administer, providing instructors with an opportunity to identify potential knowledge gaps before students get engaged in extended reading. As a result, instructors can address the knowledge gaps by pre-teaching, using other relevant interventions, or maybe even changing the reading assignment. In this way, assessment of background knowledge can help teachers adapt instruction to student needs and help students build understanding before covering a text in class.

Our approach in designing the K-tool stemmed from two core notions. First, a topical vocabulary test may become a useful instructional component, as outlined above. Second, we believe that such a tool should give teachers maximum freedom in selecting the reading passages. The teacher brings the passage they want to use in assigned reading, and the tool generates the tests. That is why the tool includes automated detection of the major topic of the passage—to allow for a wide variety of passages without prescribing supported topics in advance.

One limitation of the current system is that it is oriented for reading passages that have domain-specific topical terminology. Expository passages with STEM contents are a prominent example of such texts. Other types of reading materials can also involve special background knowledge with a prominent terminological aspect, for example, texts about specific sports like American football (Wang et al., 2021). Our initial evaluation of the K-tool was on STEM texts. We have not yet tested the tool on news or social sciences texts, in which distinct topical vocabulary is less prominent. Additional research would be needed to investigate whether the current computational approach can be extended to support texts with less prominent topical vocabularies. Another possible avenue for research is to have the tool compute automatically whether a given submitted text has as strong topical vocabulary signature. With such a component, the tool could automatically flag texts that are suitable or unsuitable for topical vocabulary testing.

Technical Limitations

In this section, we note two technical limitations of our current system that can be immediately addressed as first steps in further development: (a) handling passage titles and (b) generating the topic label for the test form.

Many reading passages used in educational settings come with titles, and in many cases, the titles are indicative of passages' main topics. However, in some cases, the passage titles are not topic-indicative (e.g., they might be catchy or humorous phrases or literary allusions). One of the obvious steps in further research would be to integrate passage titles into the topic detection process.

Another aspect of our work is yet incomplete—provision of the topic label for the test form. When the topical vocabulary test form is composed, it needs to have a topic statement, such as "Which of the following terms is related to the topic of <TopicLabel>?" For example, O'Reilly et al. (2019) presented a test form that asked, "Which of the following terms is related to the topic of ecology?" We intend for the K-tool system to generate such topic labels automatically, together with the list of test terms, for any given document.

Automatic provision of topic labels has been explored in computational linguistics literature. Bhatia et al. (2016) used word embeddings to select topic labels for topics derived from LDA processing (LDA topics are technically just sets of words). They utilized Wikipedia article titles as candidates from which topic labels could be obtained and selected the title that had the best aggregate cosine similarity with each of the words for a given topical set. The situation with K-tool is more complicated owing to several constraints. First, sometimes the suitable topic label occurs in the document and might be listed among the topical terms. Such is the case for the passage about thunderstorms (Figure 2), in which the word *thunderstorms* appears in the passage itself. In principle, it could be chosen as the query label (for posing the question) and removed from the list of terms on the test form. Yet, there is another constraint—a query label needs to be general enough, and not any specific term can be used. For example, for the thunderstorms passage, we might not want to have “hail” or “wind” selected as the query topic. The query topical label needs to be general and abstract enough to subsume all the related topical terms under it. Another requirement for the topical query label is that it needs to be easily understood by students. For example, *plate tectonics* might be a suitable general term for a text passage, but it might be unfamiliar to students (of a given grade level). Then, it would make little sense to ask students which terms are related to a topic whose label they do not understand (*earthquakes* might be more suitable in such a case). We consider that an automatic system could provide several candidate query labels to allow the teacher or researcher to choose the most appropriate one. This aspect could be addressed in future research.

Further Extensions

In this section, we describe potential extensions beyond the scope of the current prototype system. Those include: (a) use of terms beyond nouns and noun phrases, (b) consideration of text length, and (c) consideration of the number of items on a test form.

The current K-tool system generates test forms only with nouns and noun phrases. Nouns and noun phrases carry the most topical significance in a text. However, some verbs and adjectives can be distinctively topical. It is yet an open question to what extent verbs and adjectives should be integrated into topical vocabulary tests. In addition, potential inclusion of verbs and adjectives as topical terms holds the promise to alleviate the scarcity of in-document topical terms that we have encountered with some texts.

Another limitation of the current study is text length. The system showed promising generation results for reading passages of lengths from approximately 400 to 1,500 words. Passages shorter than 400 words may lack a sufficient number of topical words, or they may have just enough topical words to support only one test form. One way to address this in the future is to consider shorter test forms, requiring fewer topical terms from the document. Passages longer than 1,500 words might have enough topical terms, but they may involve more themes. The crucial aspect of such passages might be, not length per se, but their lexical diversity (how many different topical terms are used). Those aspects can be investigated in further research.

While the different levels of form difficulty are estimated via grade-level mappings, the empirical difficulty of test forms needs to be investigated in a field study with actual student populations. In addition, we need to consider the length of test forms. For the composition of a test form, we currently require 50 terms, with a breakdown of 14 in-document topical terms, 14 out-of-document topical terms, and 22 distractors. For comparison, O'Reilly et al. (2019) used test forms with 44 items, of which 26 were topical and 18 were distractors; Wang et al. (2021) used forms with 30 items. (In both studies, items were written by experts, and tests were administered to high school students.) Stahl and Bravo (2010) tested students in the second grade using forms with 25 items. Technically, we can change the required number of items or terms per form, even allowing users to control this setting, but there are open questions: Is there an optimal number of items per form? Does it depend on grade level? Maybe forms need to be longer for higher grades and shorter for lower grades. Since topical vocabulary testing is not yet a widespread educational practice, additional research will be needed to investigate those aspects and how they influence the accuracy of the vocabulary recognition test to predict students' reading comprehension. However, having an automated system to generate topical vocabulary forms can ease the creation of such tests and thus contribute to the adoption of such tests in research and in educational practice.

Considerations for Future Research

For the current K-tool implementation, we used association and similarity measures to estimate semantic relatedness between terms. While such measures provide estimates on continuous scales, for using them in the K-tool we needed to set some thresholds. The topical vocabulary test uses a yes/no response for each item. Thus, each word/term must be either obviously related to the topic or obviously unrelated. Any item that is somewhere ‘in-between’ might be considered as inconsistent (debatable) for scoring. Our major concern was that items must be scorable without much dispute. The evaluation results indicate that this requirement was met to a large extent. However, there might be some potential to use varying levels of semantic relatedness to make items more or less difficult. For example, non-topical items that are “strongly unrelated” to the topic may make a test form easier, whereas non-topical items that are “less unrelated” to the topic might make the test more difficult. Varying the levels of semantic relatedness for the topical terms can also potentially influence test form difficulty. Thus, varying the levels of semantic relatedness between terms and topics might be a good direction for future research.

Conclusion

We presented a prototype computational system, the K-tool, designed for automatic generation of vocabulary tests to evaluate whether students have the necessary background knowledge to understand content-specific reading materials. As a proof of concept, the K-tool was evaluated on texts for U.S. Grades 9 and 10. A high proportion of terms generated by the system were deemed acceptable by human evaluators. The encouraging results call for further research and development, including collecting empirical data for validation and obtaining teacher usability feedback for refinement.

Once fully established, the system will generate tests on demand, tailored to the specific content of a selected reading passage. It integrates certain lexical resources but is not dependent on any collection of educational texts or taxonomy of domains or topics. As such, it may become a useful tool for teachers and possibly for reading-content developers. Moreover, the K-tool has an integrated capability to generate test forms of different levels of difficulty by using grade-level mappings of terms.

The K-tool is generally intended for teachers’ informal use in class, as an aid to teaching, and not as a formal assessment tool. The test forms generated by the K-tool can be printed for use in class or integrated into a computer-based delivery system. They come with associated information that allows for scoring student responses automatically. The whole process of administering and scoring such forms can be very quick and thus provide a fast and efficient estimation of students’ background knowledge for reading comprehension. Students can take the topical test before a reading activity.

It should be noted that this research is the first step in a larger agenda, and many aspects of this system require additional investigation. For example, more studies need to be conducted to establish validity evidence for the use of such a system in schools.

References

Bhatia, S., Lau, J. H., & Baldwin, T. (2016). Automatic labelling of topics with neural embeddings. In Y. Matsumoto & R. Prasad (Eds.), *Proceedings of the 26th international conference on Computational linguistics: Technical papers* (COLING 2016), pp. 953–963). Association for Computational Linguistics.

Bouma, G. (2009). Normalized (pointwise) mutual information in collocation extraction. In C. Chiarcos, R. Eckart de Castilho, & M. Stede (Eds.), *From form to meaning: Processing texts automatically, Proceedings of the biennial GSCL conference 2009* (pp. 31–40). Gunter Narr.

Bullinaria, J. A., & Levy, J. P. (2012). Extracting semantic representations from word co-occurrence statistics: Stop-lists, stemming, and SVD. *Behavior Research Methods*, *44*(3), 890–907. https://doi.org/10.3758/s13428-011-0183-8

Castles, A., Rastle, K., & Nation, K. (2018). Ending the reading wars: Reading acquisition from novice to expert. *Psychological Science in the Public Interest*, *19*(1), 5–51. https://doi.org/10.1177/1529100618772271

Chau, H., Labutov, I., Thaker, K., He, D., & Brusilovsky, P. (2021). Automatic concept extraction for domain and student modeling in adaptive textbooks. *International Journal of Artificial Intelligence in Education*, *31*, 820–846. https://doi.org/10.1007/s40593-020-00207-1

Church, K., & Hanks, P. (1990). Word association norms, mutual information and lexicography. *Computational Linguistic*s, *16*(1), 22–29.

Cromley, J. G., & Azevedo, R. (2007). Testing and refining the direct and inferential mediation model of reading comprehension. *Journal of Educational Psychology*, *99*(2), 311–325. https://doi.org/10.1037/0022-0663.99.2.311

Deane, P. (2005). A nonparametric method for extraction of candidate phrasal terms. In K. Knight (Ed.), *Proceedings of the 43rd annual meeting of the Association for Computational Linguistics* (pp. 605–613). Association for Computational Linguistics. https://doi.org/10.3115/1219840.1219915

Fisher, D., & Frey, N. (2014). Content area vocabulary learning. *The Reading Teacher*, *67*(8), 594–599. https://doi.org/10.1002/trtr.1258

Flor, M., & Beigman Klebanov, B. (2014). ETS lexical associations system for the COGALEX-4 shared task. In M. Zock, R. Rapp, & C.-R. Huang (Eds.), *Proceedings of the 4th workshop on Cognitive aspects of the lexicon (CogALex)* (pp. 35–45). Association for Computational Linguistics and Dublin City University. https://doi.org/10.3115/v1/W14-4705

Flor, M., Fried, M., & Rozovskaya, A. (2019). A benchmark corpus of English misspellings and a minimally-supervised model for spelling correction. In H. Yannakoudakis, E. Kochmar, C. Leacock, N. Madnani, I. Pilán, & T. Zesch (Eds.), *Proceedings of the fourteenth workshop on Innovative use of NLP for building educational applications* (pp. 76–86). Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-4407

Flor, M., Holtzman, S., Deane, P., & Bejar, I. (2024). Mapping of American English vocabulary by grade levels. *International Journal of Applied Linguistics*, *175*(1), 25–45. https://doi.org/10.1075/itl.22025.flo

Frey, B. J., & Dueck, D. (2007). Clustering by passing messages between data points. *Science*, *315*(5814), 972–977. https://doi.org/10.1126/science.1136800

Iso, H., Wang, X., Suhara, Y., Angelidis, S., & Tan, W.-C. (2021). Convex aggregation for opinion summarization. In M.-F. Moens, X. Huang, L. Specia, & S. W.-T. Yih (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2021* (pp. 3885–3903). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-emnlp.328

Kintsch, W. (2004). The construction-integration model of text comprehension and its implications for instruction. In R. B. Ruddell & N. J. Unrau (Eds.), *Theoretical models and processes of reading* (Vol. 5, pp. 1270–1328). International Reading Association. https://doi.org/10.1598/0872075028.46

Krovetz, R., & Deane, P. (2015). Computer-implemented systems and methods for non-monotonic recognition of phrasal terms (U.S. Patent No. 9,208,145). U.S. Patent and Trademark Office.

Kurdi, G., Leo, J., Parsia, B., Sattler, U., & Al-Emari, S. (2020). A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, *30*, 121–204. https://doi.org/10.1007/s40593-019-00186-y

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159–174. https://doi.org/10.2307/2529310

Liebeskind, C., Dagan, I., & Schler, J. (2018). *Automatic thesaurus construction for modern Hebrew* In N. Calzolari et al. (Eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*. European Language Resources Association. https://aclanthology.org/L18-1229/

Lin, D. (1998). Automatic retrieval and clustering of similar words. In *36th annual meeting of the Association for Computational Linguistics and 17th international conference on Computational linguistics* (Vol. 2, pp. 768–774). Association for Computational Linguistics. https://doi.org/10.3115/980691.980696

McKeown, M. G., Deane, P. D., Scott, J. S., Krovetz, R., & Lawless, R. R. (2017). *Vocabulary assessment to support instruction: Building rich word-learning experiences*. Guilford Press.

Mulla, N., & Gharpure, P. (2023). Automatic question generation: A review of methodologies, datasets, evaluation metrics, and applications. *Progress in Artificial Intelligence*, *12*, 1–32. https://doi.org/10.1007/s13748-023-00295-9

Nagy, W., & Townsend, D. (2012). Words as tools: Learning academic vocabulary as language acquisition. *Reading Research Quarterly*, *47*(1), 91–108. https://doi.org/10.1002/RRQ.011

O'Reilly, T., Wang, Z., & Sabatini, J. (2019). How much knowledge is too little? When a lack of knowledge becomes a barrier to comprehension. *Psychological Science*, *30*(9), 1344–1351. https://doi.org/10.1177/0956797619862276

Ozuru, Y., Dempsey, K., & McNamara, D. S. (2009). Prior knowledge, reading skill, and text cohesion in the comprehension of science texts. *Learning and Instruction*, *19*(3), 228–242. https://doi.org/10.1016/j.learninstruc.2008.04.003

Pearson, P. D., Hiebert, E. H., & Kamil, M. L. (2012). Vocabulary assessment: Making do with what we have while we create the tools we need. In J. Baumann & E. Kame'enui (Eds.), *Vocabulary instruction: Research to practice* (2nd ed., pp. 231–255). Guilford Press.

Perfetti, C., & Stafura, J. (2014) Word knowledge in a theory of reading comprehension. *Scientific Studies of Reading*, *18*(1), 22–37. https://doi.org/10.1080/10888438.2013.827687

Purandare, A., & Pedersen, T. (2004). Word sense discrimination by clustering contexts in vector and similarity spaces. In *Proceedings of the eighth conference on Computational natural language learning (CoNLL-2004) at HLT-NAACL 2004*, (pp. 41–48). Association for Computational Linguistics.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-Network. In *Proceedings of the 2019 conference on Empirical methods in natural language processing* (pp. 3982–3992). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410

Sabatini, J., Wang, Z., & O'Reilly, T. (2019). Relating reading comprehension to oral reading performance in the NAEP fourth-grade special study of oral reading. *Reading Research Quarterly*, *54*(2), 253–271. https://doi.org/10.1002/rrq.226

Sannigrahi, S., van Genabith, J., & España-Bonet, C. (2023). Are the best multilingual document embeddings simply based on sentence embeddings? In A. Vlachos & I. Augenstein (Eds.), *Findings of the Association for Computational Linguistics: EACL 2023* (pp. 2306–2316). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-eacl.174

Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal*, *95*(1), 26–43. https://doi.org/10.1111/j.1540-4781.2011.01146.x

Schütze, H., & Pedersen, J. O. (1997). A cooccurrence-based thesaurus and two applications to information retrieval. *Information Processing and Management*, *33*(3), 307–318. https://doi.org/10.1016/S0306-4573(96)00068-4

Shapiro, A. M. (2004). How including prior knowledge as a subject variable may change outcomes of learning research. *American Educational Research Journal*, *41*(1), 159–189. https://doi.org/10.3102/00028312041001159

Sheehan, K. M., Kostin, I., Napolitano, D., & Flor, M. (2014). The TextEvaluator Tool: Helping teachers and test developers select texts for use in instruction and assessment. *Elementary School Journal*, *115*(2), 184–209. https://doi.org/10.1086/678294

Simonsmeier, B. A., Flaig, M., Deiglmayr, A., Schalk, L., & Schneider, M. (2021). Domain-specific prior knowledge and learning: A meta-analysis. *Educational Psychologist*, *57*(1), 31–54. https://doi.org/10.1080/00461520.2021.1939700

Smith, R., Snow, P., Serry, T., & Hammond, L. (2021) The role of background knowledge in reading comprehension: A critical review. *Reading Psychology*, *42*(3), 214–240. https://doi.org/10.1080/02702711.2021.1888348

Stahl, K. A. D. (2008). The effects of three instructional methods on the reading comprehension and content acquisition of novice readers. *Journal of Literacy Research*, *40*(3), 359–393. https://doi.org/10.1080/10862960802520594

Stahl, K. A. D., & Bravo, M. A. (2010). Contemporary classroom vocabulary assessment for content areas. *The Reading Teacher*, *63*(7), 566–578. https://doi.org/10.1598/RT.63.7.4

Sun, S., & Nenkova, A. (2019). The feasibility of embedding based automatic evaluation for single document summarization. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 conference on Empirical methods in natural language processing and the 9th international joint conference on Natural language processing (EMNLP-IJCNLP)* (pp. 1216–1221). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1116

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, *37*, 141–188. https://doi.org/10.1613/jair.2934

Vayansky, I., & Kumar, S. A. P. (2020). A review of topic modeling methods. *Information Systems*, *94*, Article 101582. https://doi.org/10.1016/j.is.2020.101582

Wang, Z., O'Reilly, T., Sabatini, J., McCarthy, K. S., & McNamara, D. S. (2021). A tale of two tests: The role of topic and general academic knowledge in traditional versus contemporary scenario-based reading. *Learning and Instruction*, *73*, Article 101462. https://doi.org/10.1016/j.learninstruc.2021.101462

Wang, Z., O'Reilly, T., & Sutherland, R. (2024). *Replicating decoding threshold in ReadBasix®: Impact on reading skills development* (Research Memorandum No. RM-24-06). ETS. https://doi.org/10.1002/ets2.12390

Wang, Z., Sabatini, J., & O'Reilly, T. (2020). When slower is faster: Time spent decoding novel words predicts better decoding and faster growth. *Scientific Studies of Reading*, *24*(5), 397–410. https://doi.org/10.1080/10888438.2019.1696347

Wang, Z., Sabatini, J., O'Reilly, T., & Weeks, J. (2019). Decoding and reading comprehension: A test of the decoding threshold hypothesis. *Journal of Educational Psychology*, *111*(3), 387–401. https://doi.org/10.1037/edu0000302

Wang, Z., Sparks, J., Walker, M. E., O'Reilly, T., & Bruce, K. (2025). Group differences across scenario-based reading assessments: Examining the effects of culturally relevant test content. In R. E. Bennett, L. Darling-Hammond, & A. Badrinarayan (Eds.), *Socioculturally responsive assessment: Implications for theory, measurement, and systems-level policy* (pp. 369–398). Routledge. https://doi.org/10.4324/9781003435105-21

Wu, X., Nguyen, T., & Luu, A. T. (2024). A survey on neural topic models: Methods, applications, and challenges. *Artificial Intelligence Review*, *57*, Article 18. https://doi.org/10.1007/s10462-023-10661-7

Xie, B., Song, J., Shao, L., Wu, S., Wei, X., Yang, B., Lin, H., Xie, J., & Su, J. (2023). From statistical methods to deep learning, automatic keyphrase prediction: A survey. *Information Processing and Management*, *60*(4), Article 103382. https://doi.org/10.1016/j.ipm.2023.103382

Appendix. Acceptance Rates for Generated Terms

Table A1 presents the final acceptance rates of terms for each of the 20 passages, with a breakdown by type of term. Note that acceptance was defined with a strict criterion, meaning the term was accepted only if both annotators marked it as accepted. The column labelled ‘Words” indicates the passage length (word count). The column “Domain” presents the general content domain of the passage.

Table A1. Acceptance Rates for Evaluated Terms, for Each Passage

Text ID	Word count	GL	Domain	TID (%)	TOD (%)	NT (%)
AP	886	9.2	Astronomy	100	79	100
DB	1,000	10.1	Biology	100	64	98
EE	804	9.7	Physics	91	82	100
FF	1,282	10.7	Biology	100	94	98
HM	1,185	9.8	Biology	100	89	100
OG	717	7.7	Astronomy	100	79	100
NSS	557	11.5	Ecology	100	75	86
SYF	1,161	13.3	Biology	100	93	100
OEA	1,229	9.8	Atmosphere	100	39	95
PC	601	9	Physics & sport	100	75	100
RWK	1,154	9.9	Physics	100	75	98
SC	827	9.9	Technology	100	75	98
SME	1,310	13.1	Biology	100	93	89
SAT	1,143	11.9	Astronomy	100	89	93
BBSS	484	7.7	Physics & sport	100	61	95
SPW	1,057	9.7	Physics & sport	100	57	100
SOB	789	7.8	Physics & sport	100	57	98
SOL	1,194	10.8	Biology	100	86	98
WHC	710	9.4	Geography	100	75	95
CFE	1,448	10.3	Ecology	100	61	100
Macro average across texts				99.5	75.4	97
Micro average across all terms				99.5	74.5	97

Note. ID = identification; GL = grade-level estimation (Flesch–Kincaid); NT = non-topical; TID = topical terms from inside the document; TOD = topical terms retrieved from outside the document.

Notes

[1] The nouns from Steps 2 and 3 would be used to generate the keys for a test form.

[2] https://opennlp.apache.org/

[3] The list of 68,000 MWEs is a subset of an even larger set of MWEs that was developed at ETS by Paul Deane and Bob Krovetz in the past. That work is yet unpublished, though it is rooted in prior work (Deane, 2005; Krovetz & Deane, 2015) that used statistical approaches over large language corpora to extract lists of MWEs. For the current work, we used only the noun MWEs.

[4] See https://en.wikipedia.org/wiki/Cosine_similarity#L2-normalized_Euclidean_distance

[5] Our current approach is a simple linear combination of cooccurrence (PNPMI) and vector similarity (cosine), weighted proportionally to the (log of) frequency of occurrence of the anchor topical terms from the document. In principle, it is possible to introduce a coefficient of importance α (e.g., $\alpha \cdot$ cosine + $(1 - \alpha) \cdot$ PNPMI) and learn the optimal value of α from empirical data. We leave such exploration for future work.

[6] The general idea is that the grade level of an expression should be higher than that of the highest individual word in it. The 25% is just a plausible heuristic. Estimating the complexity (grade level) of MWEs is not yet a developed area.

Suggested Citation:

Flor, M., Wang, Z., Deane, P., & O’Reilly, T. (2025). *Toward an automatic method for generating topical vocabulary test forms for specific reading passages* (Research Report No. RR-26-02). ETS. https://doi.org/10.64634/kb8bf328

Action Editor: Katherine Castellano

Reviewers: Mo Zhang and Jiyun Zu

www.ets.org